2020-02

Monitoring spread of epidemic diseases by using clinical data from multiple hospitals: a data warehouse approach

Rutatola, Edger Pascrates

NM-AIST

MONITORING SPREAD OF EPIDEMIC DISEASES BY USING CLINICAL DATA FROM MULTIPLE HOSPITALS: A DATA WAREHOUSE APPROACH

Edger Pascrates Rutatola

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master's in Information and Communication Science and Engineering of the Nelson Mandela African Institution of Science and Technology

Arusha, Tanzania

February, 2020

ABSRTACT

Many countries apply data science techniques to enhance their health sectors and the surveillance of diseases. The success of the innovations lies on the availability and quality of datasets to be analyzed. In Tanzania, while different Hospital Management Information Systems (HoMIS) like the Government of Tanzania Hospital Management Information System (GoT-HoMIS) are installed in various hospitals, the data stored in the systems are not integrated. This causes unavailability of high quality, timely, anonymous, harmonized, and integrated datasets that can be shared and exhaustively analyzed for epidemic diseases surveillance. This study intended to develop a data warehouse to host patients' demographic and clinical particulars essential for epidemic diseases surveillance from a multi-node GoT-HoMIS, and yield an integrated dataset that can be used for epidemic diseases surveillance.

Interviews were conducted in three strategic health facilities and the Ministry responsible for Health in Tanzania. Documents were reviewed, and observation done on the patient's registration process in the GoT-HoMIS. Thereafter, a data warehouse was developed to run under MariaDB database server, and using Hypertext Preprocessor an Extract, Transform, and Load (ETL) module was developed. The ETL module was deployed at six health facilities, and the resulting integrated dataset of 152 104 facts was visualized by using FusionCharts libraries.

The study demonstrates a novel means to extract data straight from the GoT-HoMIS nodes, which has the potential to make available and provide timely data and integrated reports for decision-making on epidemics. By scaling the innovation to other health facilities, epidemics surveillance can be significantly enhanced.

DECLARATION

I, Edger Pascrates Rutatola, do hereby declare to the Senate of The Nelson Mandela African Institution of Science and Technology that this dissertation is my own original work and that it has neither been submitted nor being concurrently submitted for degree award in any other institution.

Edger Pascrates Rutatola …… _____

Name and Signature of Candidate Date

The above declaration is confirmed by

Eng. Dr. Zaipuna O. Yonah ……. _____

Name and Signature of Supervisor Date

COPYRIGHT

CERTIFICATION

The undersigned certifies that he has read and found the dissertation acceptable by the Nelson Mandela African Institution of Science and Technology.

Eng. Dr. Zaipuna O. Yonah ……. ……..

Name and Signature of Supervisor Date

ACKNOWLEDGEMENT

First and foremost, I would like to thank the Almighty God for His grace that made all this work possible. He saw fit to bless me from the beginning to the completion of this academic pursuit in which I have learnt a lot and advanced in my career. Without His will, none of this would have been possible. And because of His will, this work will benefit my country and the world at large.

My heartfelt gratitude goes to my supervisor, Eng. Dr. Zaipuna O. Yonah. It was an honor to work with a person of your level of intellect and caliber. Your inputs and guidance throughout the research work were remarkable and made me reach for the stars that at first seemed impossible. If I retain just a fraction of what I have learnt from you and apply it in my day-to-day research activities, lucky will be those who get to work with me.

I also thank the Nelson Mandela African Institution of Science and Technology (NM-AIST) for giving me this special opportunity to pursue the Master's degree in Information and Communication Sciences and Engineering (ICSE). On the same regard, I thank Mzumbe University (MU) for granting me the study leave, and Project Two (P2) of the VLIR-UOS GRE@T programme at MU for financing the entire study. Without those two resources (time and funds), none of the work could have been done. On the same note I thank the distinguished academics in the school of Computing and Communication Sciences and Engineering (NM-AIST) for their selfless support during the entire research work.

Moreover, I thank all the people who participated in this study in one way or the other. I thank Ms. Devotha G. Nyambo for her tireless support. I thank the management of Tumbi Designated Regional Referral Hospital; the Epidemiology department of the Ministry of Health, Community Development, Gender, Elderly and Children (Tanzania); The President's Office – Regional Administration and Local Government (Tanzania); and any other entity that played part in the study.

The journey was challenging, as any other learning process. It is the love of my family who constantly cheered me, which gave me strength to get over the hurdles. I therefore thank my family for their love, support, and for believing in me. I also register my acknowledgements to my friends and my classmates as you are the ones we shared the laughter and tough times through the journey.

DEDICATION

I dedicate this work to my late parents: Pascrates A. Rutatola and Helenestina Malingumu.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS

CSS	Cascading Style Sheets
DHIS	District Health Information System
e-IDSR	Electronic Integrated Diseases Surveillance and Response
ETL	Extract, Transform, Load
GoT-HoMIS	Government of Tanzania Hospital Management Information System
HMIS	Health Management Information System
HoMIS	Hospital Management Information System
HTML	Hypertext Markup Language
ICD	International Classification of Diseases
IDS	Integrated Diseases Surveillance
IDSR	Integrated Disease Surveillance and Response
MoHCDGEC	Ministry of Health, Community Development, Gender, Elderly and Children
PHEOC	Public Health Emergency Operations Centre
PHP	Hypertext Preprocessor
PO-RALG	President's Office – Regional Administration and Local Government
TDRRH	Tumbi Designated Regional Referral Hospital
USSD	Unstructured Supplementary Service Data

CHAPTER ONE

INTRODUCTION

1.1 Background

Traditionally defined, epidemiology is a discipline that involves analysis of the distribution and determinants of health-related states or events in a specific population, and applying the acquired findings to control health problems (Szklo & Nieto, 2014). Epidemiology can either be descriptive: involving characterization of the distribution; or analytical: that is finding associations and testing hypothesis on the causes of a health-related state or event (Merrill, 2015; Szklo & Nieto, 2014). Epidemiology as a practice has a lot of advantages, one of them being the ability to identify and analyze epidemics, which generate information necessary in controlling the epidemics.

An epidemic, as defined by the New Shorter Oxford English Dictionary and quoted by Hays (2005, p. xi), "is a disease that is normally absent or infrequent in a given population but which is liable to outbreaks of greatly increased frequency and severity." Throughout the years, different epidemics have affected and caused mortality on different parts of the world and to different demographics. Some of the notable epidemics include Polio, Cholera, Tuberculosis, Epidemic Typhus/Camp fever, AIDS, Malaria, Smallpox, Yellow fever, the Black Death and the great Influenza (Edwards, 2017).

Low-income countries are greatly affected by epidemics, as they cause a significant number of overall deaths. Fig. 1 shows the ten leading causes of death in low-income countries in 2015, in which it can be observed that epidemics like Malaria, Tuberculosis and HIV/AIDS contributed most of the number of deaths.

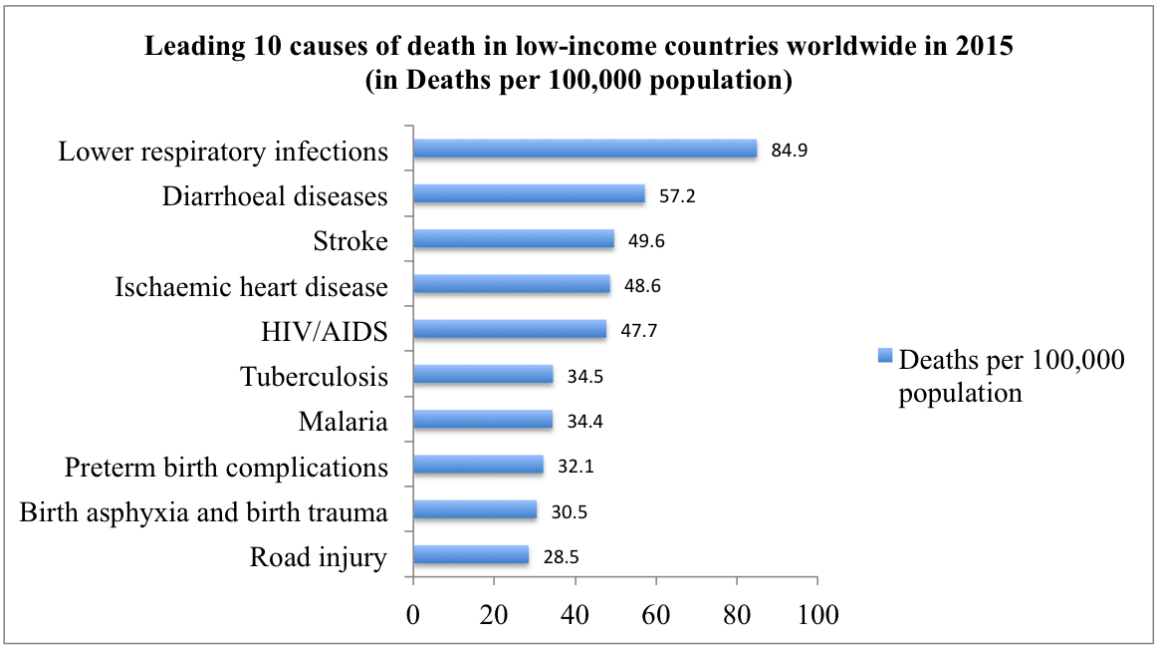


Figure 1: Leading 10 causes of death in low-income countries worldwide in 2015. Source:https://www.statista.com/statistics/311934/top-ten-causes-of-death-in-low-income-countries/

When cases of outbreaks are reported, epidemiologists can check for patterns in each case distribution and determinants so as to control the epidemics. Due to the availability of sophisticated tools and algorithms for data science and visualization, it is now relatively easier to collect, analyze and discover patterns that show the nature and characteristics of an epidemic and hence enhancing epidemiology (Koh & Tan, 2011). Different data scientists around the world work with epidemiologists to discover the patterns in catastrophes that have affected several countries in yester years. Some of the recent researches on epidemiology from past datasets include those by Tao *et al*. (2013), Shen *et al*. (2015) and Lee *et al*. (2017).

The sophisticated data mining tools and algorithms are used to extract patterns in the datasets that would otherwise be difficult, time consuming, or impossible to find (Morais, Peixoto, Coimbra, Abelha & Machado, 2017; Jovic, Brkic & Bogunovic, 2014). Availability of localized datasets is one of the powerful resources that enable developed countries to utilize data science and machine learning algorithms in epidemiology and hence having relatively advanced results and control over diseases.

On the bright side, the Government of Tanzania through its Hospital management information system (HoMIS) called Government of Tanzania Hospital Management Information System (GoT-HoMIS) has created a means to collect patients' demographic and medical diagnosis data. GoT-HoMIS is a system managed by the Tanzania President's Office – Regional Administration and Local Government (PO-RALG), which has been installed in more than 170 health facilities countrywide (PO-RALG, 2017). With the enormous amounts of data gathered in these facilities each day, big datasets can be developed that can aid the efforts of enhancing epidemiology.

However, the clinical data collected by GoT-HoMIS in the current state is multi-format and not centralized. That is, each hospital stores its own data in a local server. This is partly due to the fact that the system is still in its roll-out stage; hence the focus is on getting it installed in a large number of health facilities before focusing on centralization of clinical data storage. The progress is hindered mainly by lack of ICT infrastructure in health facilities to support installation of the system (PO-RALG, 2017). Moreover, the required network infrastructure that will enable secured centralized transfer of patients' clinical data from the GoT-HoMIS nodes in health facilities to a central server is still not set. Nevertheless, there is still hope as in some hospitals the GoT-HoMIS shares data with other remote systems such as the Government e-Payment Gateway (GePG), National Health Insurance Fund's (NHIF) system for member verification and claim, and electronic Logistics Management Information System (eLMIS) from the Medical Stores Department (PO-RALG, 2017).

In order to prepare the data collected by the individual systems for analysis and to provide meaningful and adequate results, they should be extracted from the individual databases in the hospitals, transformed (cleaned, filtered, split, merged, etc.), and loaded into a data warehouse. A data warehouse is simply a database system that collects vast amounts of data from different databases and reduces them to a form (datasets) that can be used for analytical purposes (Vaisman & Zimányi, 2014). During transformation (before being loaded into the data warehouse) the data can be filtered thoroughly to remove all the attributes that can identify a patient, for compliance to ethical code. Once the data has been loaded in the data warehouse it can be analyzed using different sophisticated data mining algorithms and tools, and even made accessible to different researchers for epidemiology purposes.

1.2 Problem Statement

The sharing of data among public health actors is hindered by many barriers. Even for the health facilities that use information systems to capture and store data, the multi-format stored data needs to be transformed to be compatible with modern analytical tools. There is also a need for technical solutions for data harmonization from different sources, integration, and sharing of the data in a way that guarantees trust among parties, without fear for misuse, abuse, or misinterpretation (Van Panhuis *et al*., 2014).

This research intended to address the problem of unavailability of high quality, timely, non-identifying (anonymous), harmonized and integrated health datasets in Tanzania to enable the applications of data science tools and algorithms to enhance surveillance of the epidemics.

1.3 Rationale of the Study

Various developed countries have applied data science to study patterns in epidemic diseases that have affected their countries in previous years, and this has proven successful. Similar research need to be done in developing countries so as to understand the characteristics of the epidemics and be able to control them. With the GoT-HoMIS being installed in more than 170 health facilities is Tanzania, it has enabled the collection and storage of patients' data when they visit the health facilities for treatment. It is therefore paramount to find a way to fetch and harmonize data from the GoT-HoMIS in a timely manner, and form an integrated dataset that can be exhaustively analysed to help understand the nature and patterns of the epidemics. Moreover, the resulting dataset should not reveal the sufferer's identity; it should be completely anonymous so that it can be shared with other stakeholders for further independent analysis.

1.4 Research Objectives

1.4.1 General Objective

The general objective of the research was to develop a software module that would extract, transform and load data from GoT-HoMIS nodes to a data warehouse to be dedicated for enhanced epidemics tracking and controlling.

1.4.2 Specific Objectives

(i) To identify and understand essential data attributes to be extracted from individual databases for transformation and loading into a data warehouse for epidemics tracking and controlling.

(ii) To design and implement the data warehouse to host the datasets extracted and transformed from the multiple sources.

(iii) To test the appropriateness of the resulting dataset for tracking and controlling spread of epidemic diseases.

1.5 Research Questions

The study was guided by seeking answers to the following research questions:

(i) What data attributes stored in the individual databases are essential in facilitating tracking and control of epidemic diseases?

(ii) How best should the data warehouse be designed and implemented to host the extracted and transformed datasets from the multiple sources?

(iii) How effective is the resulting dataset in supporting tracking and control of epidemic diseases?

1.6 Significance of the Study

Several challenges have been identified to face the Tanzanian health sector, some of them being the controlling of epidemics such as HIV/AIDS, TB and malaria; un-integrated Hospital Management Information Systems; unavailability of proper information sharing frameworks for health sector stakeholders; as well as insufficient medical informatics specialists (Ministry of Health and Social Welfare (MoHSW), 2013). Some of these challenges were confirmed by Marchant *et al*. (2014), who conducted a study in search of high quality and timely maternal and newborn health data, thus appreciating the need for data in decision making as well as international comparison and evaluations.

The importance of data for various purposes is well acknowledged, but there are still no standardized methodologies in place to ensure that the appropriate health information is channeled to the right person for right utilization (World Health Organization (WHO), 2016). Therefore, there is a need for these challenges to be addressed to upscale health information utilization, which include, among others, to improve timely compilation of data from different health information systems and various programmes, as well as building data analysis capacity.

There is no limit to the operations to enhance epidemiology that can be done on the datasets once generated, and this research focused mainly on generation of integrated health datasets from the GoT-HoMIS.

1.7 Delineation of the Study

Despite the fact that there are a number of other Hospital Management Information Systems used in different health facilities in Tanzania, this study focused on data extraction from the GoT-HoMIS nodes as a test case. This is mainly because the GoT-HoMIS has a large market share, and being government-owned, it has the potential to be installed in even more government health facilities. Therefore, it provides a large dataset to be analysed. Moreover, choosing to work with the GoT-HoMIS also influenced the study population. The participants in the study needed to have worked with the GoT-HoMIS and have a solid knowledge about it.

CHAPTER TWO

LITERATURE REVIEW

2.1 Health Information in Tanzania: The Current Status

From the report by WHO (2016), health information in Tanzania is currently generated through two major channels. The first channel consists of data at population level, which include, among others, demographic and health surveys (DHS), and HIV/AIDS and malaria indicator survey. The second channel operates routinely, and maintains two major systems, which are a Health Management Information System (HMIS) also known as MTUHA, and the Integrated Disease Surveillance and Response (IDSR). Besides the two systems in the second channel, there are also disease-specific programme systems such as those for Tuberculosis and Leprosy, Sexually transmitted diseases, as well as immunization.

There is however, a notable inadequacy of access to and utilization of the collected health information by the community as well as decision makers (WHO, 2016). This is influenced by among other reasons; the absence of standardized methodologies to ensure that appropriate information is channeled timely to the right person and for the right purpose. The recent adoption of electronic Health/Hospital management information systems will significantly improve the processes of data collection and utilization, as the data collected can be analyzed and shared easily to enhance health systems management and disease tracking and control.

2.2 Health Management Information System (HMIS)

Tanzania Government operates routine health information data system, commonly known by its Swahili abbreviation, MTUHA (Mfumo wa Taarifa za Uendeshaji wa Huduma za Afya), collecting information from more than 5400 health facilities (MoHSW, 2009). The system was conceptualized back in the early 1990s with the goal of establishing an integrated routine health data system in the country. It was implemented in 1993 to address many challenges including health facility workers being burdened with filling several forms for different reporting systems, as well as the vastness of the data being collected with little capacity for analysis and utilization at all levels (MoHSW, 2009). It is further praised to have evolved to become the key tool for

monitoring and evaluating health sector reform performance using indicators to measure the performance with respect to the targets (MoHSW, 2009).

The same report by the MoHSW acknowledges that the biggest strength of the HMIS is its coverage, in that it is implemented across the whole country at health facilities, hence collecting vast amount of data. Nonetheless, it has some shortcomings such as the presence of gaps and overlaps in the collected datasets for reporting; poor data quality; no integrated framework to handle and analyze all the data; infrequent integrated analysis across data sources; needed information not reaching the targeted stakeholders such as the regional managers; the resulting information showing only aggregated data; and lastly but not least, the use of paper based registers (manual process) for data collection and reporting (MoHSW, 2009; Mandara *et al*., 2005).

2.3 District Health Information Software

The District Health Information Software (DHIS2) is free and open source software for collection, validation, reporting, analysis and presentation of data to support health managers. The data and indicators through DHIS2 are analyzed and presented using different tools such as pivot tables, standardized reports and thematic maps (MoHSW, 2009; Mandara *et al*., 2005). In the same report by the MoHSW, it can be deduced that despite its strength in analyzing data, the system is configured to operate in stand-alone mode without direct communication with the systems at district and regional offices, and hence the data is not available in a timely manner. Additionally, data quality audits conducted still indicate suboptimal quality of the contained data (Ministry of Health, Community Development, Gender, Elderly and Children (MoHCDGEC), 2016). At present, aggregated clinical data from all health facilities in Tanzania are integrated and stored in the DHIS2 on a monthly basis, where analytics are performed and reports are generated. The DHIS2 contains vast amount of data that is effectively analyzed to provide reports, making it the main system from which epidemic diseases surveillance is done. Data are fed in the DHIS2 at the district level in aggregated manner monthly, after being collected from individual facilities (Darcy, Somi, Matee, Wengaa & Perera, 2017; MoHCDGEC, 2016). For more enhanced analytics and study of patterns in disease cases, the data needs to be at individual level (dispersed) rather than aggregated. Aggregation of the data can be as part of the reports

generated from it, but the ability to analyze the data at individual case level is key to epidemic diseases surveillance. Moreover, epidemics surveillance needs to be timely and comprehensive so that actions towards an epidemic are effective (Health Metrics Network (HMN), 2008).

2.4 Electronic Integrated Disease Surveillance and Response

The electronic Integrated Disease Surveillance and Response (eISDR) system is one among the data sources to the DHIS2. It goes a step further to enable the usage of mobile phones to submit real-time notifications about disease cases as well as weekly reports. The eIDSR utilizes Unstructured Supplementary Service Data (USSD) services to send these reports to the targeted recipients who are mainly epidemiologists. The system by itself, without the mobile phone component, dates back to 1998, which was adopted to detect and assist in responding to epidemics (Rumisha, Mboera, Senkoro, Gueye & Mmbuji, 2007). From the eISDR, weekly and monthly reports are submitted to the District Medical Office, which can then be accessed through the DHIS2 by epidemiologists for surveillance of epidemic diseases.

2.5 Hospital Management Information Systems

Increase in number of local software developers and availability of open source software has led to improved development and management of various Hospital Management Information Systems (HoMISs) in developing countries (Karuri, Waiganjo, Daniel & Manya, 2014; PO-RALG, 2017). In the URT, a number of Hospitals (public and private) operate different Information Systems for storage and manipulation of clinical and administrative information, which include the GoT-HoMIS, Jeeva and Care2x, among others (PO-RALG, 2017; Nyasulu, Kasubi, Boniface & Murray, 2014; Wambura, Machuve & Nykänen., 2017). Adoption of the Information systems is made possible due to the support and emphasis from the Government on development and operationalization of such systems. Notable advantages have been gained from the systems, including but not limited to generation of reports that help in hospital administration as well as monitoring clinical operations.

These systems are not centralized, even at the minimum level for those of the same platform. The collected data are therefore only limited in the facility on which they operate. Furthermore, at the

current state, there is no integration amongst the systems to enable sharing of patient information once a patient attends different facilities for treatment. Each facility maintains its own isolated data of the patient. This type of setting is not very beneficial when it comes to epidemic diseases surveillance. The data from the HoMISs needs to be centralized so as integrated reports can be generated from a comprehensive dataset, enhancing tracking and control of the epidemics. Due to the heterogeneity of the systems, creating a data warehouse is a good approach where the data can be timely extracted from the source databases in the host facilities, transformed accordingly, and loaded in the data warehouse. Through running data science and machine learning tools and algorithms on the resulting dataset, surveillance of the epidemics as well as availability of informative reports for decision-making will be enhanced.

Despite the adoption of these HoMISs by different facilities, currently none is integrated directly to other successive systems for joint analytics such as the DHIS2. Therefore, the hospitals running these systems still have to draft monthly aggregated data, which are then fed into the DHIS2 for generation of integrated reports. From the aggregated data, only limited analytics and patterns can be observed.

2.6 GoT–HoMIS

The GoT-HoMIS was developed initially to solve the problem of "pipeline leakage" in user-fees collection at the Tumbi Designated Regional Referral Hospital (TDRRH) (KEC, 2015). It started as a system with four modules, namely, the patient registration, revenue collection, pharmacy management, and exemption management. It had tremendous impact on boosting revenue collection at TDRRH from 300 000 Tanzanian Shillings (Tshs) to 3 000 000 Tshs per day. From this remarkable success, the system was installed at different hospitals and its adoption has been successful. Some examples being like in Sekoutoure and Songea Regional Referral Hospitals whose daily collection rose from 900 000 Tshs to 2 500 000 Tshs, and 280 000 Tshs to 1 350 000 Tshs, respectively (KEC, 2015).

The system then evolved to become a comprehensive web and modular based system having more than 25 hospital management modules, ranging from patient care services to the overall hospital management, generating more than 38 different reports. The strength of GoT-HoMIS lies on the fact that it is modular, scalable, developed by local specialists, and managed by the

President's Office – Regional Administration and Local Government (Tanzania) (Kibaha Education Centre (KEC), 2015). It is currently installed in over 170 health facilities, but the PO-RALG has a vision to install it in all regional, district hospitals, and other health facilities (PO-RALG, 2017). The development of the system is ongoing, with a dedicated team at PO-RALG receiving feedback from users and strengthening the system's functionalities. While such a wide adoption of the system could be advantageous for data mining and analytics, the GoT-HoMIS nodes are currently not centralized and hence analysis of collected data is only limited to the host facility.

2.7 Data warehouse applications and design

2.7.1 Similar data warehouse applications

A data warehouse refers to a special database designed to support decision making. It consolidates data from various sources and transforms it to new structure that is best suited for decision support. The sources of data to a data warehouse can be operational databases or any other systems internal or external to a particular organization (Vaisman & Zimányi, 2014). The application of data warehouses to extract data from multiple sources for enhanced decision support is not new, even in the Tanzania health sector. The current operational data warehouse in the Tanzania health sector is the DHIS2 described in Section 2.3. The data warehouse has also been installed in several other African countries (Sæbø, Kossi, Titlestad, Tohouri & Braa, 2011).

Another example is the application of a data warehouse at Mayo Clinic (United States of America) to integrate data from patient care, education, research, and administrative transactional systems, and organizing them to support business intelligence and decision making. As a result, among other benefits, the data warehouse at Mayo clinic has enabled a single standardized reporting and analysis environment for infection data (Chute, Beck, Fisk & Mohr, 2010). Also in the United States of America, another data warehouse called the Virtual Data Warehouse (VDW) is used by the Health Care Systems Research Network (formerly known as the HMO Research Network). The first principle and primary objective of the VDW is to facilitate research in public domain health and health services. The data in the VDW can be categorized into seven (7) content areas: enrollment/demographics, utilization, laboratory, pharmacy, census, tumor registry, and vital signs/social history. In 2014, the VDW captured data

from 17 sites, which collectively contributed to over 185 million person-years of data to be used for research by scientists and research staff (Ross *et al*., 2014). A dataset of such magnitude being available for research can lead to profound discoveries and knowledge that can improve the health sector.

2.7.2 Dimensional data modeling

From definition and primary purpose of a data warehouse, it serves to collect data from various sources and stores them in a way that enables information retrieval and decision making. The design of a data warehouse is therefore influenced by the information needs of the targeted users of the resulting datasets and the way the data are organized in the source systems (Imhoff, Galemmo & Geiger, 2003). Most of the source data structures are designed for operational (transactional) needs. Therefore, extra effort is needed in determining which attributes are to be extracted and how they are to be stored in an analytical system (data warehouse). The two most popular approaches to modeling data from source databases to a data warehouse are Entity-Relational and dimensional modeling (Sen & Sinha, 2005). Entity-Relational modeling follows the standard online transactional processing database design, which comprises of entity-relationship (ER) design (conceptual schema), forming the relational schema, and normalizing the relational schema. On the other hand, a dimensional model is composed of a fact table, joined to a number of dimension tables (ibid.).

In simple terms, dimensions of a data warehouse are the key descriptors based on which facts can be organized and analyzed, whereas, a fact is a single instance of a particular occurrence (Hammergren & Simon, 2009). Technically, a fact table is a special table that comprises of a composite key, formed from surrogate keys of several dimension tables (ibid.). According to Kimball and Ross (2013), dimensional modeling is a more preferred way of modeling analytic data as it delivers data that is understandable to targeted users and leads to optimal query performance.

In dimensional modeling, each dimension table of a data warehouse has a one-attribute key that corresponds to one attribute of the composite key of a fact table, resulting into a multiple dimensions to a single fact table. This forms a design resembling a star, and hence it is called star schema (ibid.). Figure 2 is an example of a star schema.

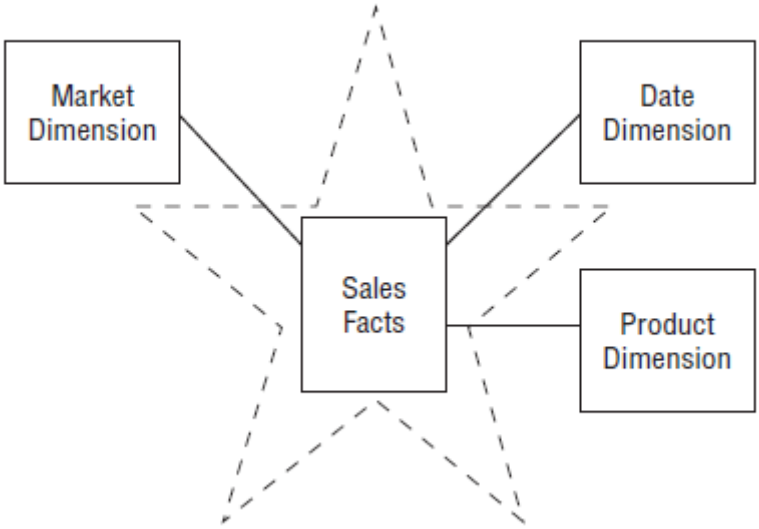


Figure 2: An example of a star schema (Kimball and Ross, 2013).

However, due to the design of the data source and nature of the data, some of the dimension tables may need to be linked to each other as a means of getting linked to the fact table. Such a design is called snowflake schema (Cherniack, Lawande & Tran, 2014).

2.8 Conclusion and research gap

After examining information systems innovation in the Tanzania health sector, it is concluded that there is still a need for timely extraction of data from the varying HoMISs operated at different health facilities for joint analysis, especially of epidemic disease cases. Commendable efforts have been done on utilization and evolution of the DHIS2, but it still struggles in terms of timeliness and completeness of the contained data. Therefore, this study set out to find a means of timely extraction of data directly from the GoT-HoMIS nodes in the health facilities, harmonize them, and load them into a well-designed data warehouse to form a unified dataset that can be exhaustively analyzed to enhance epidemic diseases surveillance. Globally, data warehouses have been used for similar goals, but they need to be properly designed to yield both meaningful reports and optimal performance.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Study Population

In this study, involved population comprised of medical doctors, epidemiologists, health information systems focal persons, and health information systems administrators. These were either based at the Ministry of Health, Community Development, Gender, Elderly, and Children (MoHCDGEC), or among three strategic health facilities located in Morogoro, Dar es Salaam, and Pwani Regions in the United Republic of Tanzania. The chosen health facilities were Tumbi Designated Regional Referral Hospital (TDRRH) (Pwani Region - Tanzania), Mzumbe Health Centre (Morogoro Region - Tanzania), and the Muhimbili University of Health and Allied Sciences (MUHAS) (Dar es Salaam Region - Tanzania). These were purposively selected based on their roles and position in the Tanzanian health sector (MoHCDGEC); experience in operating HoMISs, specifically the GoT-HoMIS (TDRRH and Mzumbe Health Centre); and having a nominated HoMIS pioneer (MUHAS).

A total of ten (10) representative respondents from the mentioned facilities were involved in the study. These were purposively obtained through snowball sampling, based on their key roles and knowledge on epidemiology, and experience in working with the GoT-HoMIS or other Health Management Information systems and innovations in the country. Snowball sampling (network sampling) was adopted as the researcher needed to interview the lead people in health systems innovations in the Tanzanian health sector, hence each respondent was asked to nominate the next eligible candidate until saturation point was reached. With exception of the experts who had worked with or pioneered the GoT-HoMIS, the rest of the focal people for other health information systems were not known to the researcher, thus snowball sampling was fit for the study (Naderifar, Goli & Ghaljaie, 2017; Polit & Beck, 2010). Seven (7) of the ten respondents were medical doctors and/or epidemiologists, who were further divided into two categories. The first category is of those who are stationed in hospitals. This category had four (4) medical doctors; one being stationed at Mzumbe Health Centre, two at TDRRH, and one from MUHAS. These were chosen following their experience in working with HoMISs, particularly the GoT-HoMIS. The second category (Team) comprised of doctors and epidemiologists stationed at the MoHCDGEC in Dar es Salaam, Tanzania. This category had three (3) doctors/epidemiologists;

one being the national IDSR focal person, another one being the Public Health Emergency Operations Centre (PHEOC) manager, and lastly a full-time epidemiologist. This category was selected based on their key knowledge on epidemiology and their roles in the current systems and innovations in support of the same.

The remaining three (3) respondents were not of the medicine or epidemiological background. One was a statistician at the MoHCDGEC and a national DHIS2 focal person. The other two were systems administrators having more than four years of experience working with the GoT-HoMIS.

3.2 Data collection

Semi-structured interviews were conducted on March and April 2018 to obtain primary data. The interviews intended to identify key attributes to be extracted from the GoT-HoMIS nodes in the health facilities that can help in facilitating timely monitoring and analysis of epidemic cases. The other part of the interviews, specifically dedicated to respondents from the MoHCDGEC, aimed to derive information on the current setup for reporting on disease cases, especially epidemics, from when the diagnosis is made at a health facility to the district, regional, and national responsible teams.

Observation method was also employed on the patients' registration processes at the Tumbi Designated Regional Referral Hospital. Firsthand observation was carried out on how patients were registered over time. The objective was to examine the quality of the data fed into the GoT-HoMIS in terms of correctness and completeness.

In addition, documents were also collected and reviewed to obtain additional knowledge. Review was done on literature and various documents focusing on health information systems and the setup for reporting and monitoring in Tanzania's health sector. Some of the documents were obtained online, whereas some were provided by focal persons at the MoHCDGEC.

3.3 Guiding model

Cross-Industry Standard Process for Data Mining (CRISP-DM) was adopted as a guiding model for the project. It involves six phases that are conducted in a cyclic manner. The six phases are:

business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Koh & Tan, 2011; Morais *et al*., 2017). CRISP-DM was selected as the guiding model due to its several internal feedback loops between the phases, leading to consistent and reliable results (Kabakchieva, 2013). It is also easier to use and widely used in data mining projects (Munirathinam & Ramadoss, 2016).

The current GoT-HoMIS operations and the entire reporting setup from when a disease case is identified in a health facility to when it is reported at the district, regional, and national level were clearly examined in the business understanding phase. The development of the data warehouse and ETL modules to host a comprehensive yet timely dataset was identified as an objective. In the data understanding phase, the data contained in the GoT-HoMIS nodes was well studied from point of entry to the point of storage. This was done through observation at TDRRH, and later followed by deep study of the GoT-HoMIS schema and close discussions with the GoT-HoMIS development team. During the data preparation phase, the data warehouse was designed and developed, along with the custom ETL algorithms. The ETL scripts were run on the GoT-HoMIS nodes fetching data from the nodes and loading it to the developed data warehouse. Upon simulation of some analytics on the resulting integrated datasets and studying the results during the modeling phase, they were constantly evaluated and the team went back to the data preparation phase, analyzing the ETL module over and over until it produced a desired dataset.

3.4 Data warehouse design and development

Literature was reviewed to guide the design and development of the data warehouse. The GoT-HoMIS database schema was studied exhaustively as it is the main data source and the design of the data warehouse depends on how the needed data are mapped in it. The data warehouse was designed and developed to run under MariaDB database environment. This was primarily motivated by the fact that the source database is designed for and runs on the same, and hence there will be no compatibility issues. Moreover, MariaDB is more improved in terms of ease of use, performance, and bugs fixes as compared to MySQL, making it a better option (Bartholomew, 2012).

3.5 Extraction, transformation, and loading

Hypertext Pre-processor (PHP) was used as a scripting language for development of algorithms that would extract the data from the GoT-HoMIS nodes (source databases), transform them accordingly, and load them in the data warehouse. This was after thoroughly studying the GoT-HoMIS database schema and working with the GoT-HoMIS developers to understand the relationships and characteristics of the contained data. Hypertext Pre-processor was selected based on the fact that it is open source; easy to learn and implement to develop web-based applications; it has been used to develop powerful systems such as Facebook, Wordpress, Yahoo!, and Wikipedia; and it also works well with MariaDB database engine, which is the engine that runs the GoT-HoMIS database (Walia & Gill, 2014).

3.6 Testing the resulting dataset

Once the ETL module was run, the resulting dataset had to be tested for its merits towards epidemic diseases surveillance. The module was run on GoT-HoMIS databases at Tumbi Designated Regional Referral Hospital (TDRRH), Mount Meru Regional Hospital, Bukoba Regional Referral Hospital, Sumbawanga Regional Referral Hospital, Biharamulo Desiganted District Hospital, and Rubya Hospital. The hospitals were purposively selected based on the criteria of having full-fledged GoT-HoMIS operational for at least three months by the time of testing. Once the integrated dataset from the facilities was obtained, FusionCharts library was used to visualize various reports emanating from the dataset. FuisonCharts library is a JavaScript library that makes it easier to produce numerous types of charts, with the ability to be integrated with various programming languages including PHP. It is free to use for non-commercial uses and paid-for if the uses are commercial (Supaartagorn, 2016). It also has available documentations and free help across the web.

A user interface was created to aid selection of parameters for reports to be produced as well displaying the reports. The interface was developed by using Hypertext Markup Language (HTML), Cascading Style Sheets (CSS) and JavaScript. Hypertext Pre-processor was used as a backend scripting language to process the user inputs and rendering the desired report. The resulting dataset was frequently evaluated in collaboration with liaisons from the GoT-HoMIS team and the ETL algorithms were constantly perfected.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Respondents' Profiles

Seven (7) of the ten respondents (medical doctors and/or epidemiologists) from the MoHCDGEC, TDRRH, and Mzumbe Health Centre had academic qualifications and working experiences as presented in the charts of Figs. 3 and 4:

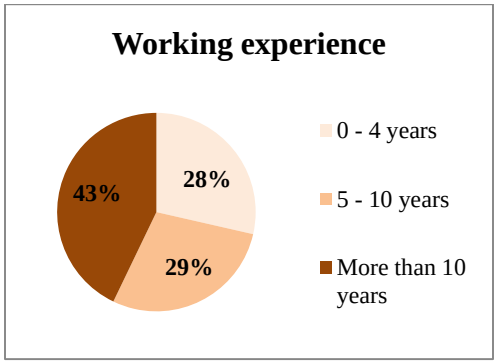


Figure 3: Respondents' working experience.

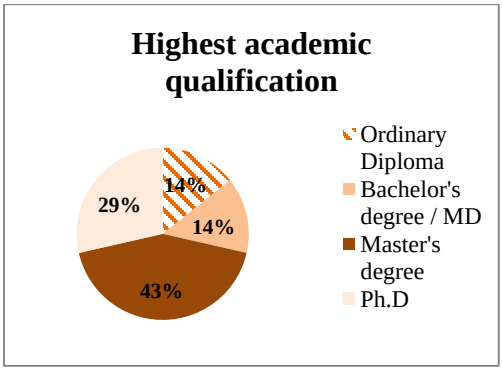


Figure 4: Respondents' highest academic qualification.

The level of education and working experience of the respondents were sought so as to determine their expertise on their line of work, particularly dealing with epidemic diseases surveillance (epidemiologists) and usage of information systems (medical doctors working with GoT-HoMIS, and focal people for other information systems). From Fig. 3, it can be observed that 72% of the respondents had more than five years of working experience in their line of work, while 43% had more than ten years of experience. Moreover, Fig. 4 shows that more than 72% of the respondents had Master's degree or PhD as their highest level of educational qualification. These findings show that the involved respondents in the study were indeed experts in their fields of work.

4.2 Needed Patients' Particulars for Enhanced Epidemiology

During interviews, respondents (with exception of the system administrators) were provided with a list of patients' particulars collected by the GoT-HoMIS when a patient is registered at the Hospital for the first time. The particulars included first name, middle name, surname, gender, date of birth, age, mobile number, tribe, residence, marital status, occupation, country, and next of kin (name, residence, relationship, phone number). The respondents were asked to select particulars out of the list, and suggest any other particulars (demographic and/or clinical) that can be used to trace patterns in disease cases and eventually enhance epidemiology and epidemics surveillance (Appendices 1 - 3). The goal of this exercise was to determine the particulars to be captured from the GoT-HoMIS databases in health facilities to the data warehouse, and the information to be contained in the analytical reports that would be produced from the data warehouse to enhance epidemic diseases surveillance. These would guide the design and development of the data warehouse and algorithms to produce desired reports. Figure 5 and Fig. 6 are screenshots of the GoT-HoMIS registration form at TDRRH.

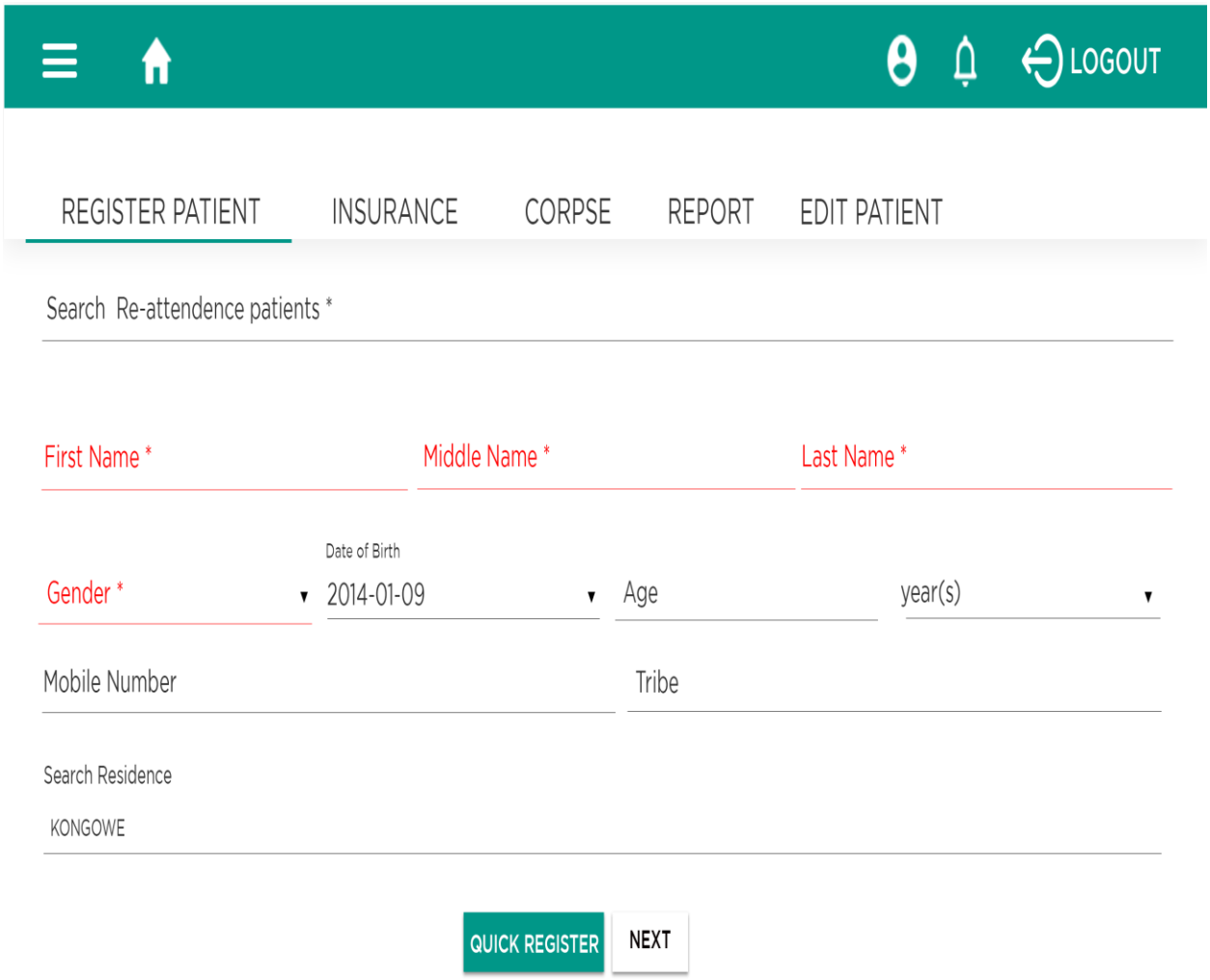


Figure 5: GoT-HoMIS patient registration form.

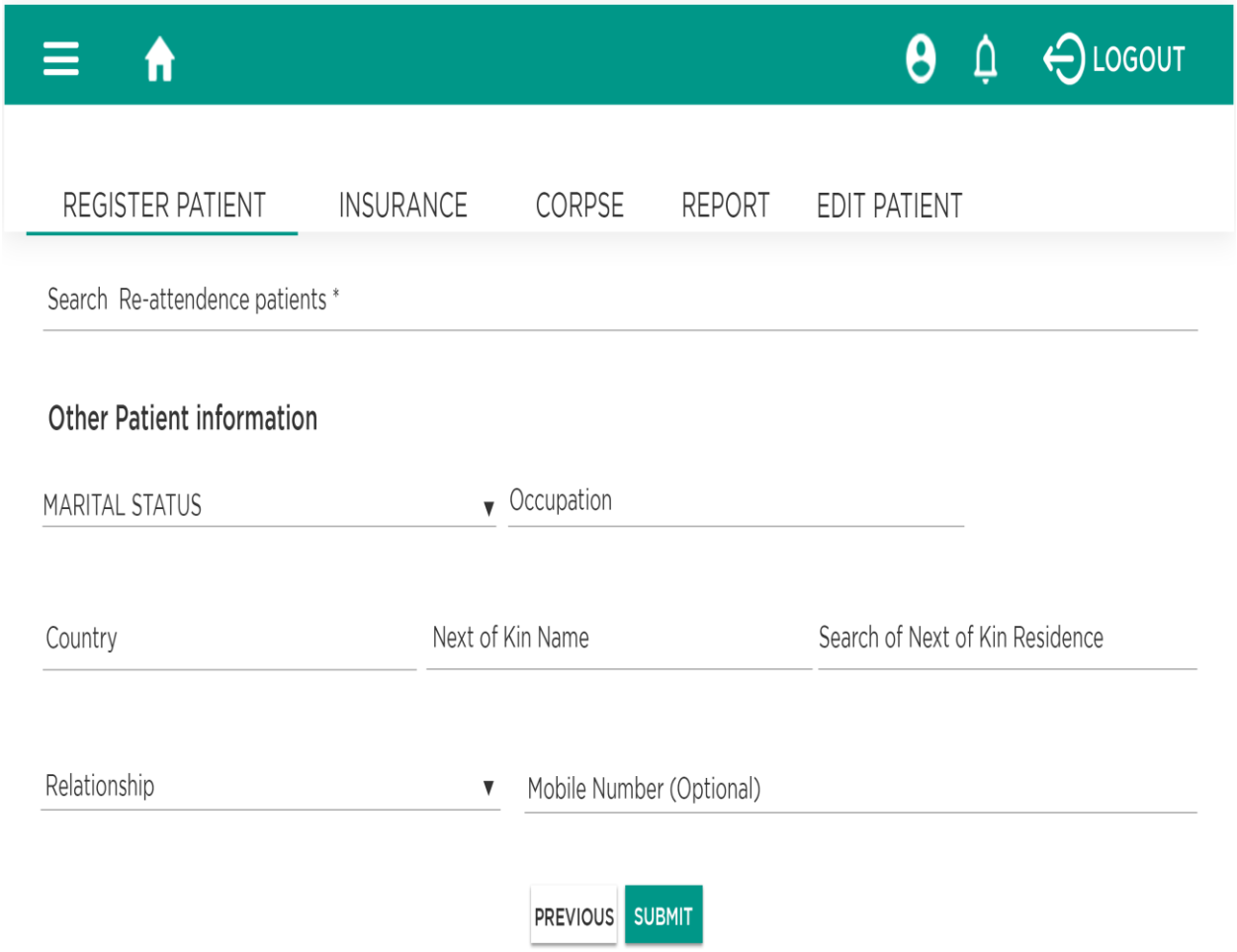


Figure 6: GoT-HoMIS patient registration form.

Out of the list, 100% (n = 7) of the respondents recommended the inclusion of gender, age, and residence as parameters in epidemics surveillance. Furthermore, 85.71% (n = 6) of the respondents voted for occupation, 57.14% (n = 4) for diagnosis, and 42.86% (n = 3) for both tribe and marital status. In parallel, 14.29% (n = 1) suggested religion despite it not being in the list. Those in support of tribe and religion associated them with some cultural practices that may lead to transmission of diseases. Figure 7 is a visualization of the suggested attributes along with the number of respondents in support of each:

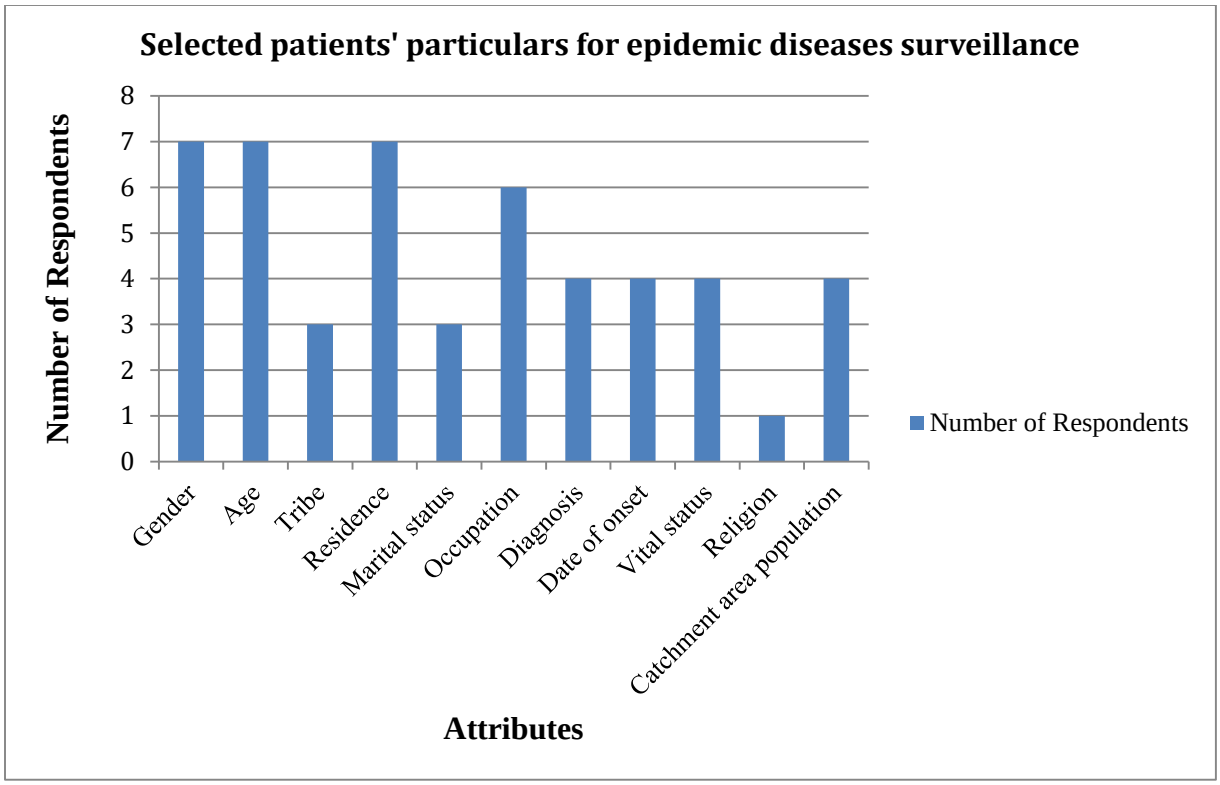


Figure 7: Selected particulars from the GoT-HoMIS for epidemics surveillance.

Catchment area population and vital status, though not included in the list provided to the respondents, were each proposed by four of the respondents (57.14%). They further explained that the two attributes would be useful to provide insights on the severity of the case and enhance informed decision making on mitigation strategies. The catchment area population and the vital status would give insights to the attack rate (AT) and case fatality rate (CFR), respectively. For example, reporting fifty (50) cases of a disease in a population of two hundred (200) people is quite different from fifty (50) cases in a population of five thousand (5000) people. Moreover, five cases of a disease being confirmed at a hospital and all five-people dying within a short time is different from five cases of a disease where all or a significant number of the affected are still alive past the first week. Vital status was also closely associated with date of onset, which also was proposed by 57.14% (n = 4) of the respondents.

4.3 The Current Reporting Setup at the MoHCDGEC

The respondents at the MoHCDGEC were further interviewed on the current reporting process from when a disease case is confirmed at a health facility to the district, regional, and national level, and diseases surveillance in general. The target was to understand the current setup and the role of the existing systems and innovations in facilitating timeliness, completeness, and enhancement of the reports. With the information flow and associated technologies in place properly known, it is easier to build a solution that complements the existing ones in achieving a common goal. The current setup derived from interviews responses is as presented in Fig. 8:

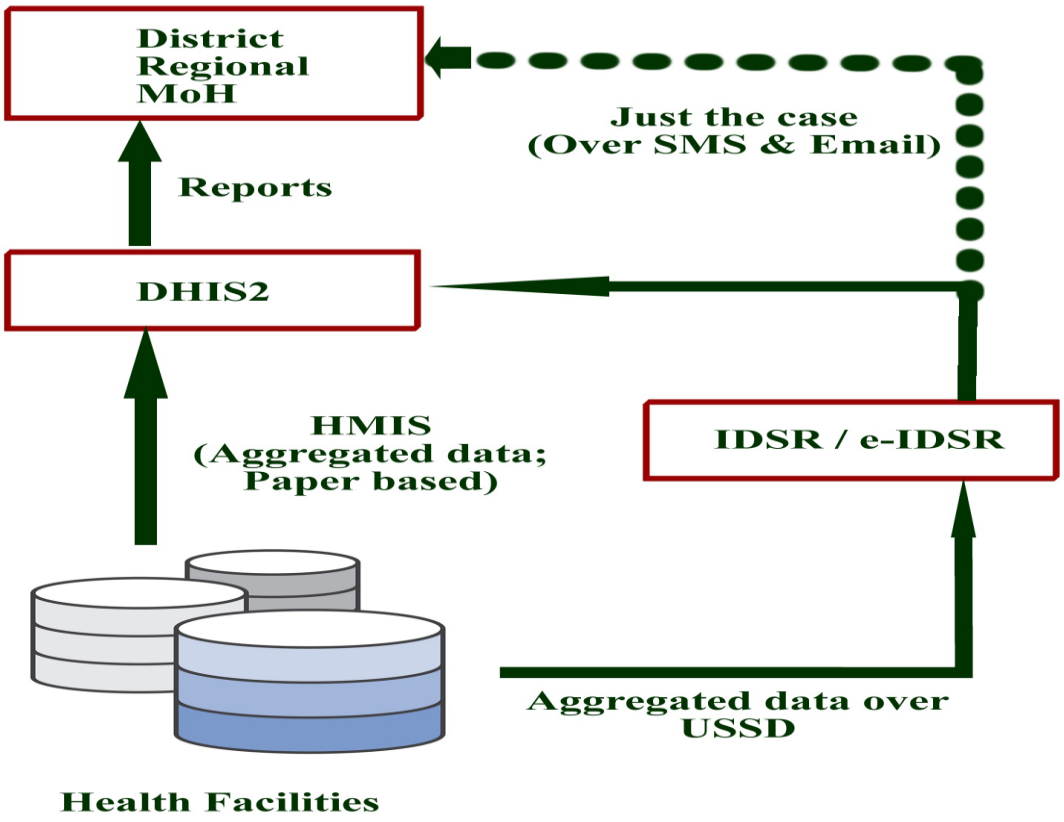


Figure 8: Current setup for disease cases reporting (Rutatola, Yonah, Nyambo, Mchau & Musabila, 2018).

It was revealed in the interviews, as presented in Fig. 8, that the health facilities are the primary sources of information. These capture information about patients and their visits at the facilities.

Different hospitals operate different Hospital Management Information Systems such as the GoT-HoMIS, Jeeva and Care2x; while, others are still paper based. Harmonization of the collected data is done through the Health Management Information System (HMIS/MTUHA), paper-based registers that contain aggregated reports on different aspects regarding clinical matters. There are total of 16 registers, each having its own distinctive purpose and information. The patients' clinical data from the health facilities through HMIS reports are inserted in the District Health Information System (DHIS2). This operation is done on a monthly basis. Moreover, the data inserted in the DHIS2 is already in aggregated manner. The DHIS2 generates various reports that are circulated to the district, regional, and national health teams. Despite the quality of reports generated by the DHIS2, nature of the primary data entry system (from hard copy registers) and being on monthly basis results in challenges of timeliness, correctness, and completeness of reports. Some disease cases require immediate alert to epidemiologists and time-to-time follow-up, which is impossible with the DHIS2. In addition to the shortcomings, it was reported that there are often mismatches between the aggregated data presented by the DHIS2 and the data present in the health facilities.

The respondents further explained that in response to the need for timeliness of reporting and alerts for some diseases, especially outbreaks, the MoHCDGEC developed the Integrated Disease Surveillance and Response (IDSR) and its more enhanced innovation e-IDSR. These were developed to simultaneously feed data to the DHIS2 and alert district, regional, and national responsible personnel in case of an outbreak. The e-IDSR utilizes Unstructured Supplementary Service Data (USSD) to pass information to the DHIS2 as well as targeted epidemiologists. Selected people at the health facilities provide information to the system, and alerts are immediately forwarded to the epidemiologists in forms of emails and text messages. The alerts however, contain just the case (diagnosed disease), the name of the health facility where the diagnosis has been confirmed, and phone number of the contact person at the health facility for the sake of follow-up. Figure 9 shows a sample text message sent to epidemiologists.

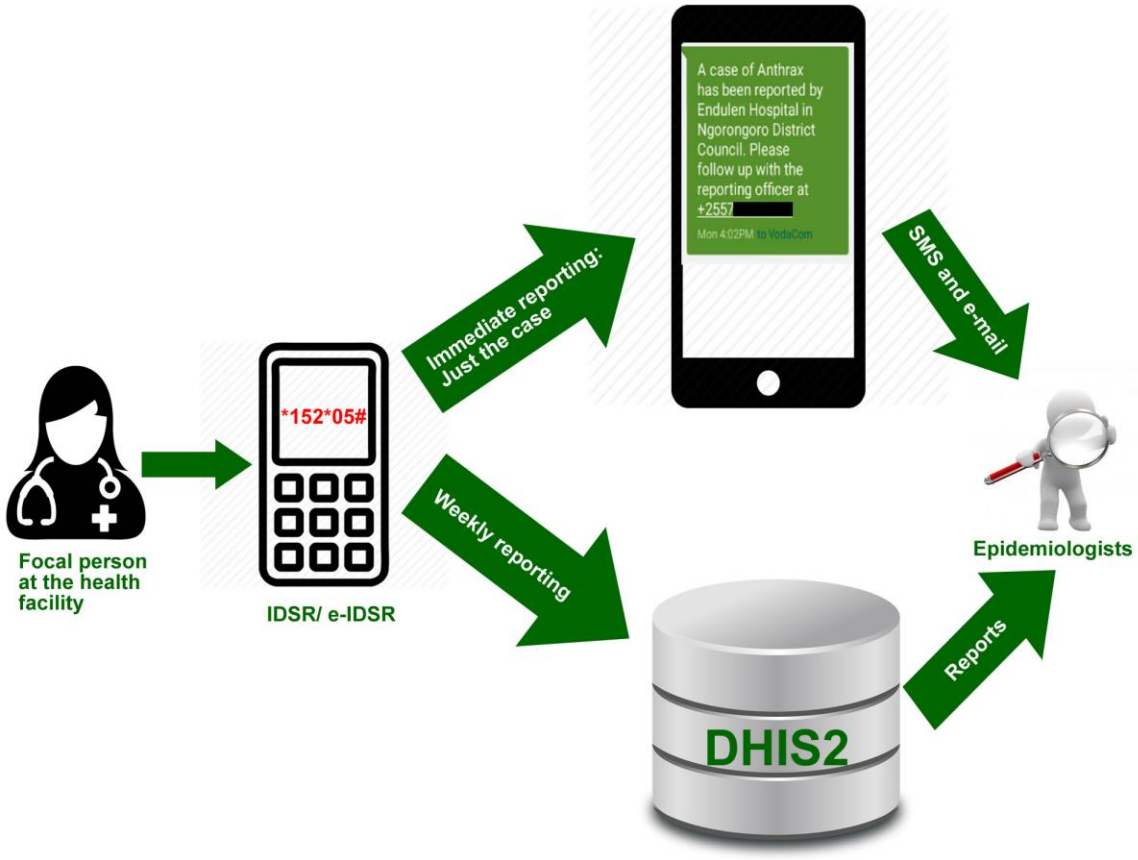


Figure 9: e-IDSR framework (MoHCDGEC, 2018; Rutatola *et al*., 2018).

The epidemiologists also pointed out that this has proven to be insufficient information to work with in terms of analysis and looking for patterns in the outbreak. More details about the characteristics of the disease case would have been more helpful. On a few selected diseases (epidemic prone diseases), weekly follow-up is done by feeding data into the IDSR (in aggregated manner), which consequently feeds them into the DHIS2. Reports can then be generated from the DHIS2 grouping the aggregated number of cases gender and age wise as seen in the sample report format in Fig. 10.

FORM 3 C: WEEKLY REPORTED NEW CASES / DEATHS DURING AN EPIDEMIC AT REGION LEVEL

Region: Week beginning: Week ending:

S/N	DISEASES	< 5				> 5			
		C		D		C		D	
		M	F	M	F	M	F	M	F
1	AFP								
2	Anthrax								
3	Blood Diarrhea								
4	Cholera								
5	CSM								
6	Human Influenza/SARI								
7	Keratoconjuctivitis								
8	Measles								

Figure 10: Template for an aggregated report generated by the DHIS2 (MoHCDGEC, 2018).

One of the challenges facing e-IDSR usage is the USSD not being efficient in times of outbreaks and in large hospitals where there is a large number of disease cases. USSD are also subject to timeouts when there is delay in the feeding process (Lakshmi, Gupta & Ranjan, 2017). While the timeouts in USSDs are implemented as a security feature, they are less friendly in the hospital setup where a data entrant may be slow or inexperienced rather than trying to tamper with the request. Moreover, it was documented that the MoHCDGEC must pay a monthly fee for the aggregator that collects and aggregates data from the USSDs and feeds them to the DHIS2 (MoHCDGEC, 2018).

The e-IDSR is currently operational in 15 regions in the URT. Furthermore, three (3) respondents pointed out that data and reports generated from the health facilities only reflect a fraction of health issues and disease cases present in the country. Disease cases in the community may not be captured unless the sufferer reports to a health facility, which is not always the case. The respondents further requested for a way that community data (disease cases) can be captured and integrated in the reports for better understanding and planning of diseases mitigation strategies. Moreover, the epidemiologists mentioned rumors as one of the main sources of their

information of outbreaks. Due to the advancement of technology, internet connectivity, and smartphones ownership, most rumors are aired on social media platforms. It was further stated that a significant number of the received rumors prove to be valid, and they reach the epidemiologists through their connections quicker that reports from hospitals.

4.4 Proposed Data Integration Framework

The findings of the survey pointed out a need for a data integration framework that would guarantee inclusion of data from various sources (identified patients' particulars from the GoT-HoMIS and related data from other HoMIS in the country, social media, meteorological data). For this reason, a framework has been envisioned to guide the integration of data from the diversified sources, joint analytics, and timely as well as detailed presentation of reports to enhance epidemic diseases surveillance and informed decision-making. The framework targets prompt analytics and reporting to support the Public Health Emergency Control Centers (PHOEC) as emphasized by the World Health Organization (2016). Figure 11 illustrates the proposed Data Integration Framework.

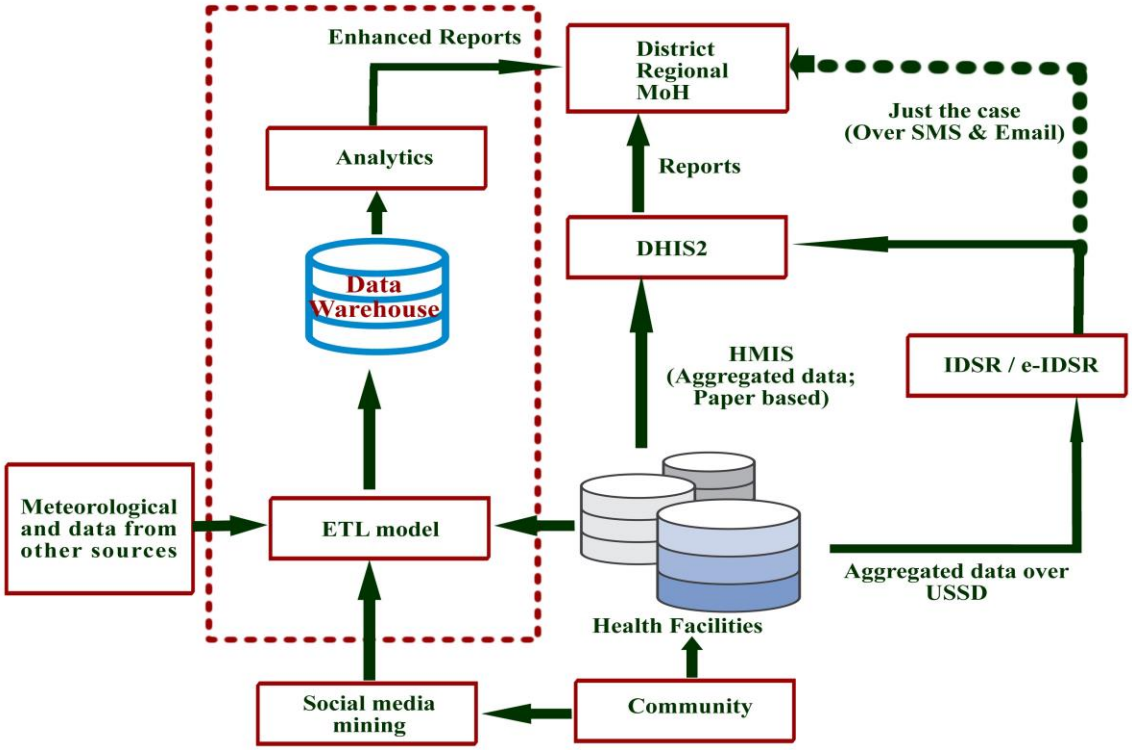


Figure 11: Data Integration Framework (Rutatola *et al*., 2018).

The data integration framework complements the existing reporting setup illustrated in Fig. 8. The extraction, transformation, and loading (ETL) model is targeted to fetch data from the HoMISs in the health facilities (GoT-HoMIS, Jeeva, Care2x), transform them accordingly and load them into a data warehouse for joint analytics.

To enhance analytics and epidemic diseases surveillance, the ETL model will also extract and transform data from social media and load it to the data warehouse. To some extent, inclusion of means to jointly analyze rumors and epidemic diseases related posts from social media will mitigate the problem of missing analysis of community data for sufferers who do not visit health facilities. Social media have proven to be a strong source of information about disease cases and outbreaks. A good example of a case where the spread of an epidemic was widely posted on social media is the case of Ebola, in which countless posts were being shared on social media (Fernández-Luque & Bau, 2015; Househ, 2016). Once the posts that contain targeted phrases or mere mention of epidemics have been extracted, they can be forwarded to epidemiologists timely, and be utilized for epidemic diseases surveillance. If well designed, social media mining can be a great contribution to epidemic diseases surveillance as well as other health related issues (Fernández-Luque & Bau, 2015; Nikfarjam, Sarker, O'connor, Ginn & Gonzalez, 2015; Paul *et al*., 2016). Prompt alerts to the epidemiological teams based on timely extraction of rumors from social media will also increase effectiveness in decision making and controlling the case(s). Addition of meteorological data for integrated analysis will enable observation of patterns of disease cases in relation to the weather or climate changes.

All the aforementioned joint analytics are missing in the existing setup at the MoHCDGEC. The diversification of data sources will increase the possibility of patterns observation and extensive analytics, and consequently enhance epidemic diseases surveillance. On the resulting dataset, superior to the existing, data mining and machine learning algorithms can be applied to produce comprehensive reports and prediction of outbreaks for effective control of the epidemics. The framework does not obliterate the existing setup and innovations; the e-IDSR should still send alerts to the epidemiological teams at the district, regional, and national levels. The alerts, however, should be linked to the associated comprehensive reports that are generated from analysis on the integrated dataset.

Following the adoption of HoMISs in a notable fraction of health facilities, extracting the required demographic and clinical data straight from the databases in the facilities will lead to timely, complete, and relatively more accurate data as compared to the current setup. Therefore, a new module needs to be added in the HoMISs to extract the required data from the databases for timely integrated analytics. The data loaded in the proposed warehouse should be transformed (on individual patient-wise, not aggregated) to allow extreme data mining operations. Aggregation will be done within the data warehouse as part of analysis should the need arise, and not before. Extraction of raw data from the sources will also reduce the mismatch in reports, contrary to what is currently experienced by the DHIS2 from the reports present at the health facilities. In case of an outbreak, the epidemiologists can still receive a notification as they do now from the e-IDSR, but through the proposed framework they can find timely system-generated analysis and determined patterns among new and historic cases.

4.5 Data Warehouse Design and Development

The data warehouse was designed following the snowflake schema. This was concluded as necessary after studying how needed data to be extracted are originally mapped in the GoT-HoMIS database structure, which led to the need for some dimensions to be linked to each other as a way to be linked to the fact table. Attributes identified during the interview were adopted as dimensions of the data warehouse. The data warehouse was therefore designed to have twelve (12) dimensions namely; region, council, facility, residence, tribe, occupation, marital status, patient, date, diagnosis, diagnosis type, and gender dimensions. Connecting them all is a fact table, comprising of surrogate keys from the dimensions as foreign keys. The data warehouse schema is as seen in Fig. 12.

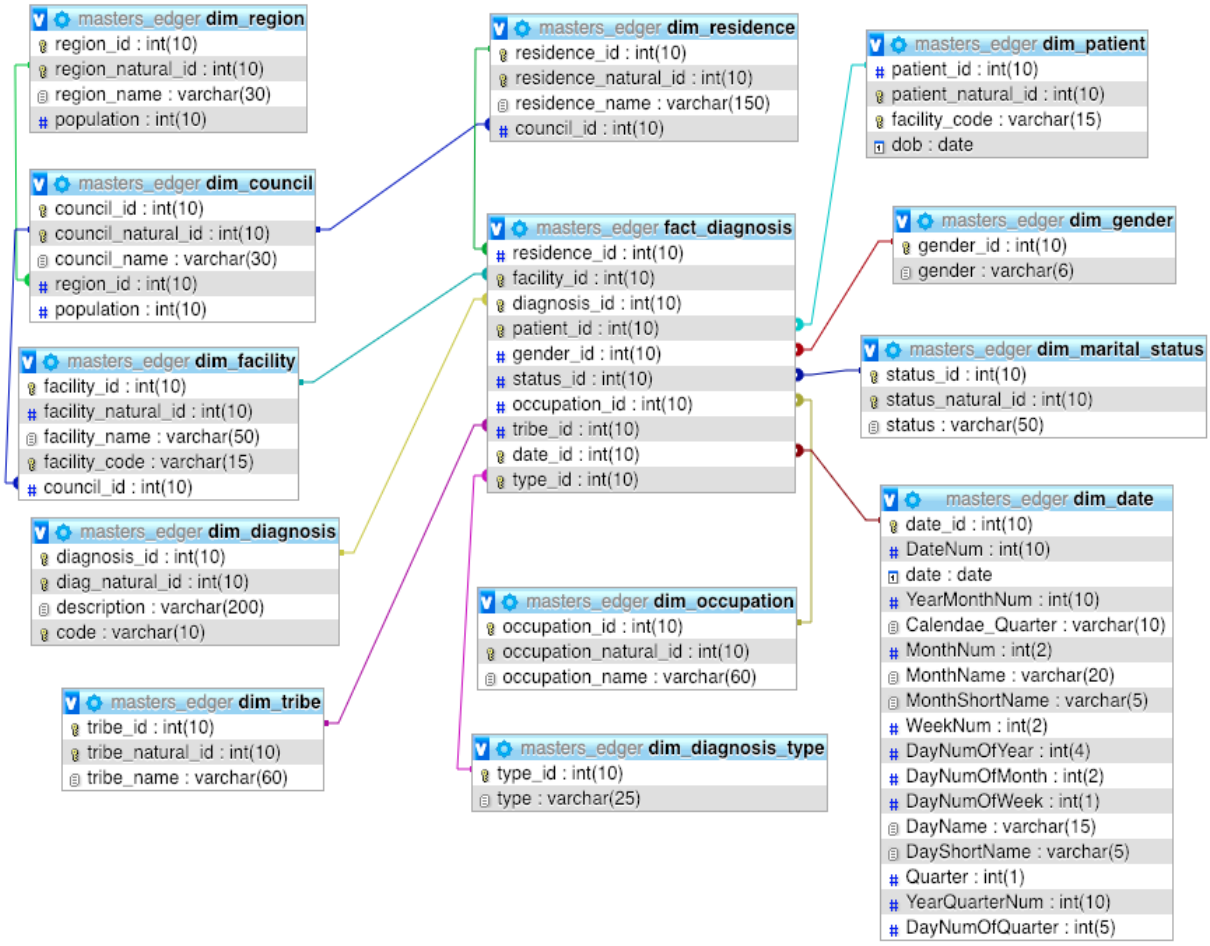


Figure 12: The data warehouse design.

The tables with prefix “dim_” are dimension tables, whereas, that with prefix “fact_” is a fact table. Dim_region and dim_council are linked, which then get linked with the fact_diagnosis through the dim_residence and dim_facility tables (snowflake schema). The major role of the region and council dimensions were to contribute to the analytics with respect to catchment area population. They can also be used in grouping of the facilities and disease cases for council-wise or region-wise analytics. All the other dimensions are directly linked to the fact tables as can be seen in the Fig. 12. For clarification, the diagnosis dimension (dim_diagnosis) is designed to host all the diagnosis descriptions along with their codes as defined in the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10). The

diagnosis type dimension (dim_diagnosis_type), on the other hand, is designed to host the types of diagnosis, which are provisional, confirmed, or differential.

It can also be observed in the design that all dimension tables but the dim_date, dim_diagnosis_type, and dim_gender have two sets of "id" columns. The natural_id refers to their primary key columns as defined in the GoT-HoMIS (source) database schema. They are defined here with the "unique" constraint to avoid duplication of loaded records. The dimensions in the data warehouse have been assigned auto-incrementing integers as surrogate keys. These are the ones used in all references within the data warehouse. Kimball and Ross (2013) also recommend this design approach. The natural_id attribute links the records in the data warehouse to those in the source tables within the GoT-HoMIS.

The date dimension was also included to facilitate analytics categorized into different time periods and seasons. It also adds the potential of integrating the data warehouse with meteorological data for improved analysis, as illustrated in Fig.11. Being characterized with attributes like quarters of the year and months, can easily be integrated with associated weather data and hence provide more insights. Pre-loading the date dimension with data and linking it with associated climatic conditions can also be useful in the analysis of disease patterns and surveillance. It can also be helpful in prediction of future outbreaks of a disease if it is somehow associated with the weather.

The data warehouse was then developed to run in the MariaDB database environment. phpMyAdmin tool was used to assist the development of the data warehouse through its graphical user interface. The database server version on which the data warehouse was set to run faultlessly was "mysql Ver 15.1 Distrib 10.1.37-MariaDB."

4.6 Extraction, Transformation, and Loading

A custom ETL module was developed to fetch data from the GoT-HoMIS nodes. The GoT-HoMIS database schema, which contains 250 normalized tables was studied thoroughly. The GoT-HoMIS developers were also consulted to clarify some areas that were not clear at first glance. The tables hosting the needed particulars were identified and the attributes' data types noted. Hypertext Preprocessor (PHP) was used as the server-side scripting language to fetch and

manipulate data from the nodes to the data warehouse. The developed module was named Integrated Diseases Surveillance (IDS).

The first script file was developed with the main objective of fetching data from the GoT-HoMIS nodes to the dimension tables. Two PHP-MySQL connections were established. The first connection was to the source database, while the second connection was to the data warehouse. Region data were first loaded into the regions dimension (dim_region). This was given priority as the region_natural_id is needed as foreign key in the council dimension (dim_council), as seen in Fig. 12. The next instructions were to extract data from the council table in the GoT-HoMIS to the council dimension, with region_natural_id from the regions dimension appearing as a foreign key (region_id) carrying the references to regions as it is on the GoT-HoMIS schema. The facility and residence dimensions were populated next. These two had foreign keys referencing the council dimension (dim_council). The instructions for populating the rest of the dimension tables with exceptions of the date, diagnosis type, and sex dimension then followed. Figure 13 is a flowchart of the IDS algorithm that was used to extract data from the GoT-HoMIS tables to the dimension tables of the data warehouse.

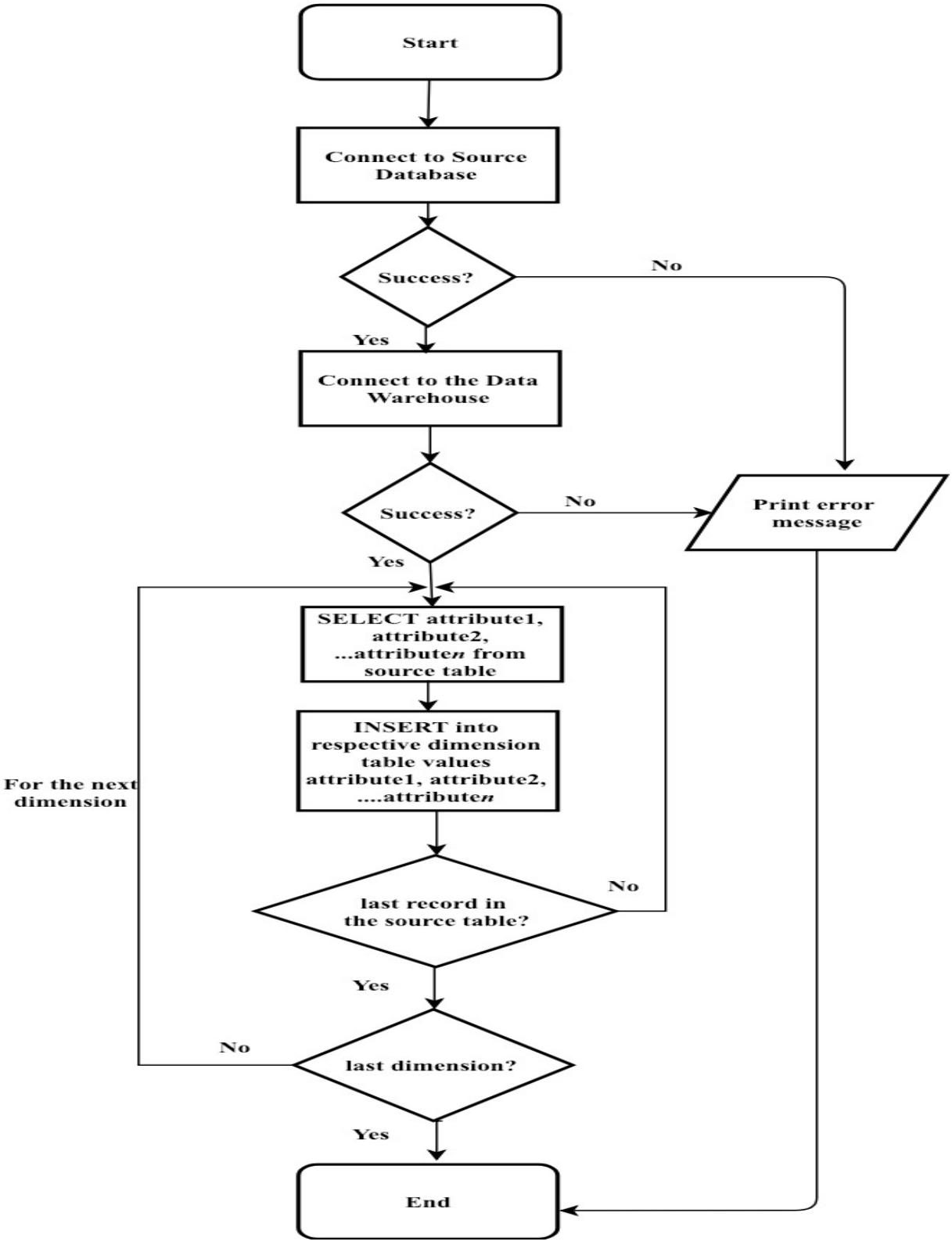


Figure 13: Flowchart for IDS algorithm; extracting data from the GoT-HoMIS to dimension tables.

The gender dimension (dim_gender) was populated through the phpMyAdmin platform, as it required only two values, Male and Female. The diagnosis type (dim_diagnosis_type) too, as it included static values provisional, confirmed, and differential. The date dimension (dim_date) was developed and populated to host the dates along with their descriptions from January 1, 1991 to December 31, 2039 as freely provided to by Sisense (Tommy, 2017). The dates were loaded from a comma-separated values (CSV) file into the dim_date.

In the dimension tables, a new entry was inserted containing "NULL" in the describing attribute. This was designed to be the identity for all data that has not been filled in the source database, thus having NULL for the dimension values. Kimball and Ross (2013) also recommend this approach.

Once all the dimension tables had been filled with data and the developed script tested to update the dimension tables with new data, a script to load data in the fact table was developed (the second script file). The fact table consists of a composite key formed by the surrogate keys of the patient (patient_id), diagnosis (diagnosis_id), diagnosis type (type_id), facility (facility_id), and date (date_id) dimensions, as designed in Fig. 12. The selection of the attributes for the composite key was based on an argument that "the same patient (patient_id) cannot be received in the same health facility (facility_id), be diagnosed with the same sickness (diagnosis_id), the diagnosis being of the same type (type_id for either differential, confirmed, or provisional), on the same day (date_id), more than once." Other attributes in the fact table are residence_id, status_id, gender_id, occupation_id and tribe_id from the residence (dim_residence), marital status (dim_marital_status), gender (dim_gender), occupation (dim_occupation), and tribe (dim_tribe) dimensions, respectively. The script was then run to populate the fact table (fact_diagnosis) with corresponding data from the dimension tables, as they relate to the patients' diagnosis records from the GoT-HoMIS database. Figure 14 is a flowchart of the IDS algorithm that extracts the needed attributes from the GoT-HoMIS database to the fact table.

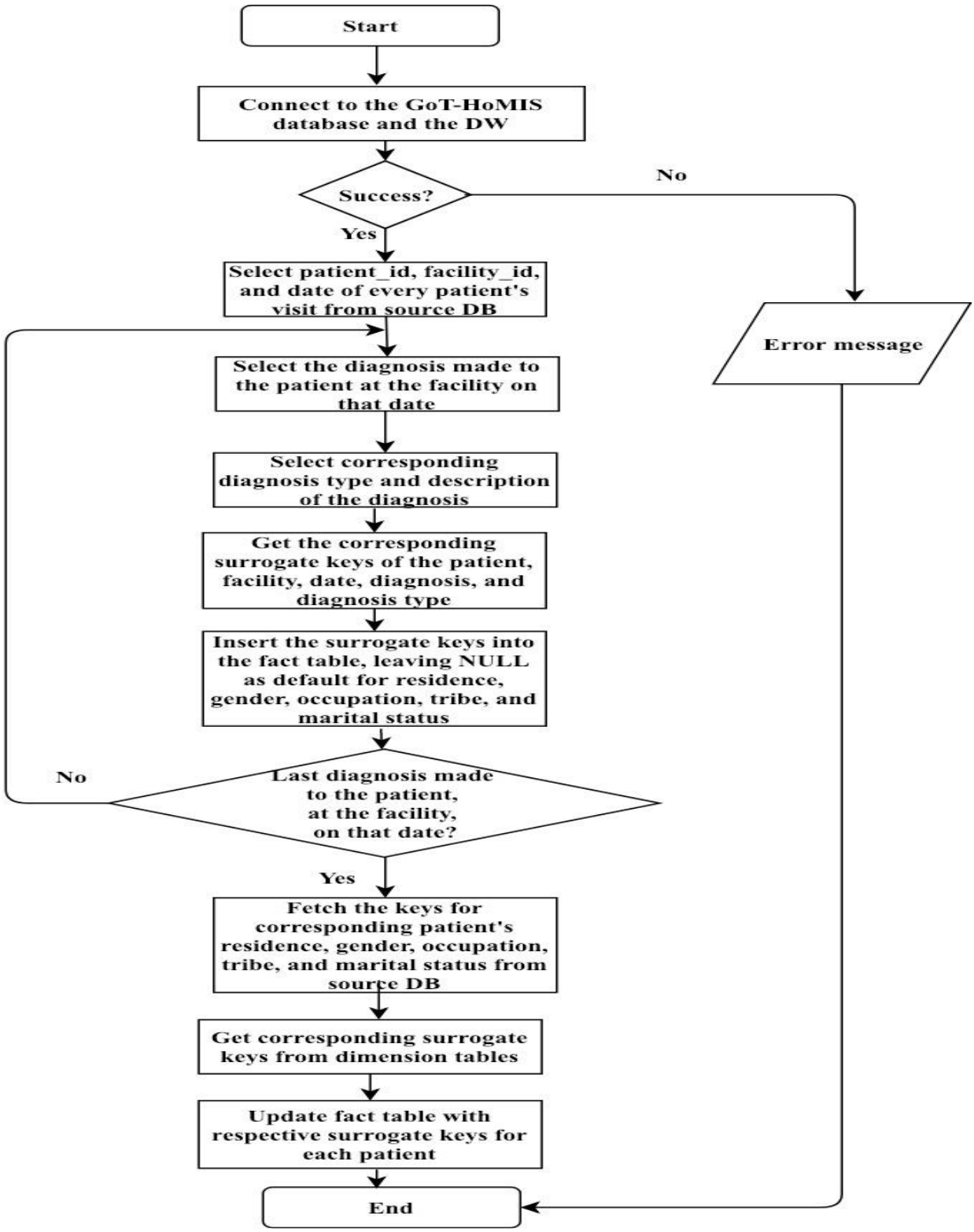
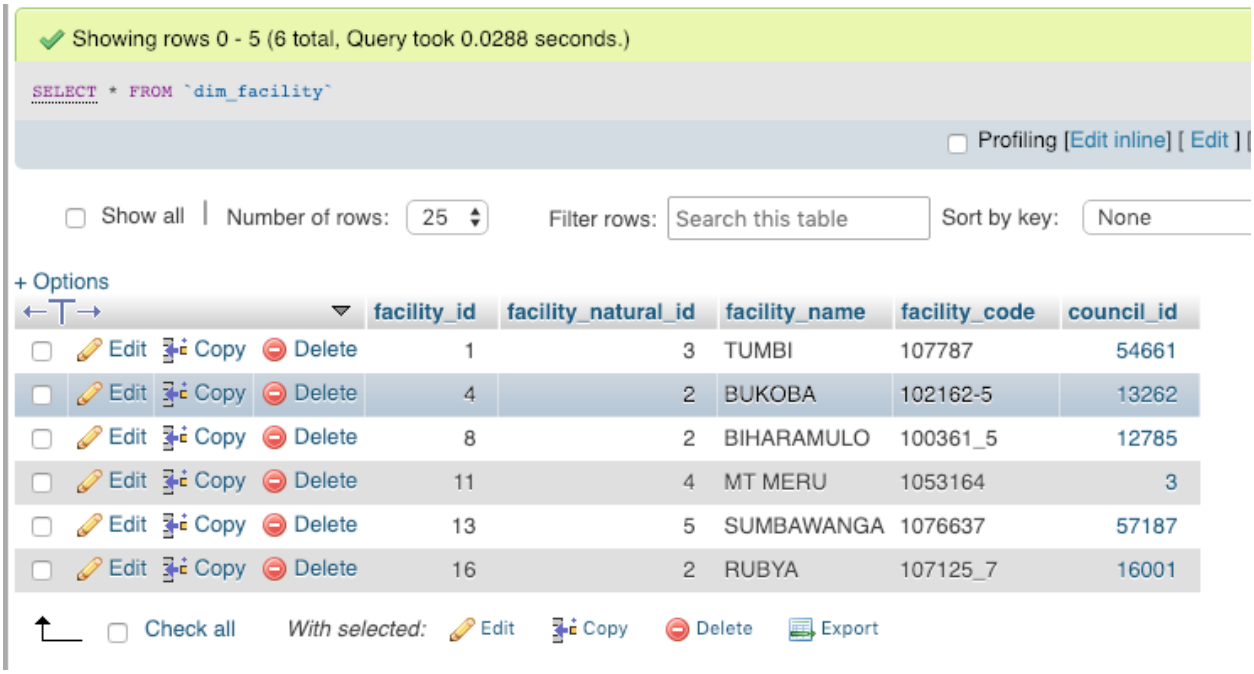


Figure 14: Flowchart for IDS; extracting data from the GoT-HoMIS database to the fact table.

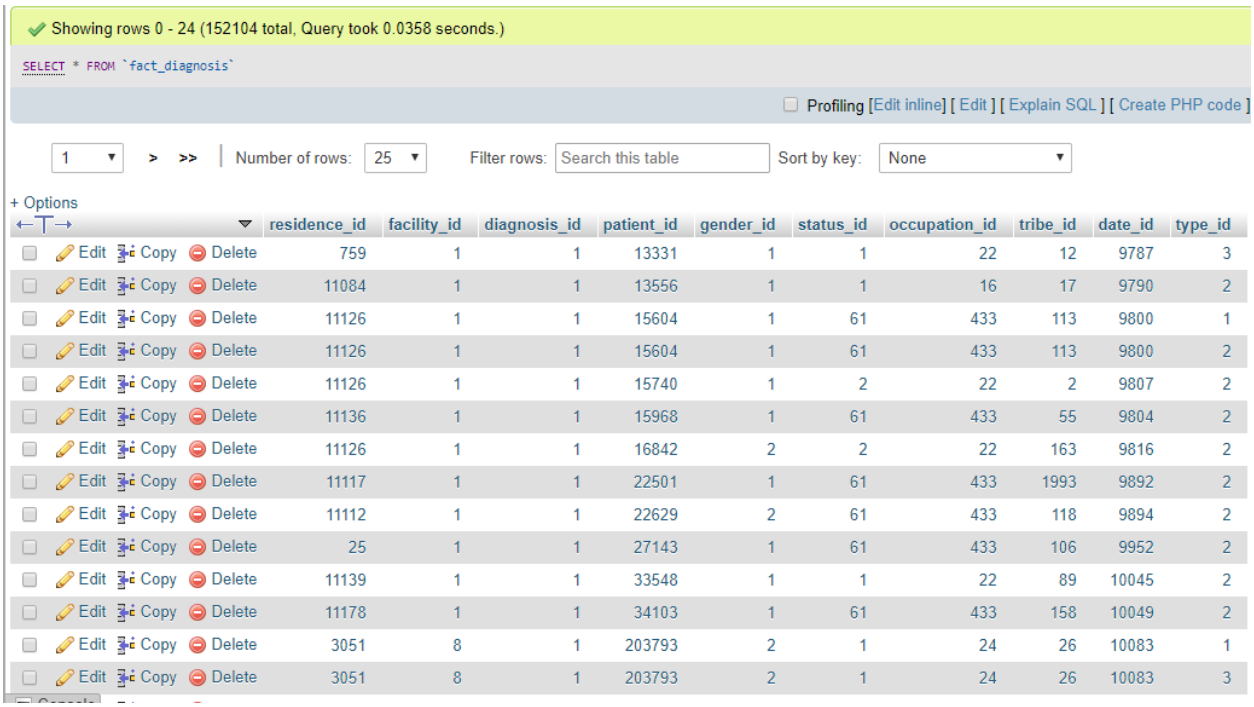
4.7 Testing the Resulting Dataset

For testing, the developed scripts were run on six (6) GoT-HoMIS database nodes, which are: Tumbi Designated Regional Referral Hospital (TDRRH), Mount Meru Regional Hospital, Bukoba Regional Referral Hospital, Sumbawanga Regional Referral Hospital, Biharamulo Designated District Hospital, and Rubya Hospital. The scripts successfully populated the dimension and fact tables. As a result, 283 379 patients information were extracted from the GoT-HoMIS in the six health facilities to the data warehouse, specifically in the patients' dimension (dim_patient), and 152 104 facts were generated in the fact table (fact_diagnosis), on which analytics are to be made and reports developed. Figure 15 and Fig. 16 show snapshots of populated facility dimension and the fact table, respectively.



	facility_id	facility_natural_id	facility_name	facility_code	council_id
Edit Copy Delete	1	3	TUMBI	107787	54661
Edit Copy Delete	4	2	BUKOBA	102162-5	13262
Edit Copy Delete	8	2	BIHARAMULO	100361_5	12785
Edit Copy Delete	11	4	MT MERU	1053164	3
Edit Copy Delete	13	5	SUMBAWANGA	1076637	57187
Edit Copy Delete	16	2	RUBYA	107125_7	16001

Figure 15: List of facilities populated in facility dimension (dim_facility).



	residence_id	facility_id	diagnosis_id	patient_id	gender_id	status_id	occupation_id	tribe_id	date_id	type_id
Edit Copy Delete	759	1	1	13331	1	1	22	12	9787	3
Edit Copy Delete	11084	1	1	13556	1	1	16	17	9790	2
Edit Copy Delete	11126	1	1	15604	1	61	433	113	9800	1
Edit Copy Delete	11126	1	1	15604	1	61	433	113	9800	2
Edit Copy Delete	11126	1	1	15740	1	2	22	2	9807	2
Edit Copy Delete	11136	1	1	15968	1	61	433	55	9804	2
Edit Copy Delete	11126	1	1	16842	2	2	22	163	9816	2
Edit Copy Delete	11117	1	1	22501	1	61	433	1993	9892	2
Edit Copy Delete	11112	1	1	22629	2	61	433	118	9894	2
Edit Copy Delete	25	1	1	27143	1	61	433	106	9952	2
Edit Copy Delete	11139	1	1	33548	1	1	22	89	10045	2
Edit Copy Delete	11178	1	1	34103	1	61	433	158	10049	2
Edit Copy Delete	3051	8	1	203793	2	1	24	26	10083	1
Edit Copy Delete	3051	8	1	203793	2	1	24	26	10083	3

Figure 16: Populated fact table (fact_diagnosis).

The resulting dataset was visualized by using FusionCharts library, well integrated within PHP scripts. A user interface was developed to aid the process of retrieving and displaying various reports originating from the hosted dataset. The user interface contained checkboxes for a user to select the health facilities whose data should be included in the generation of an integrated report. A user can check one or more options, and only data from the chosen facilities will be jointly analyzed. Furthermore, a user needs to specify a disease case about which they want to view reports, followed by a dropdown menu containing the parameters that the user can use to analyze the disease case (suggested parameters from the interview responses, presented in Fig. 7), and the time periods on which the generated report should base. Figure 17 through Fig. 21 show the designed user interface.

Figure 17: The homepage of the user interface.

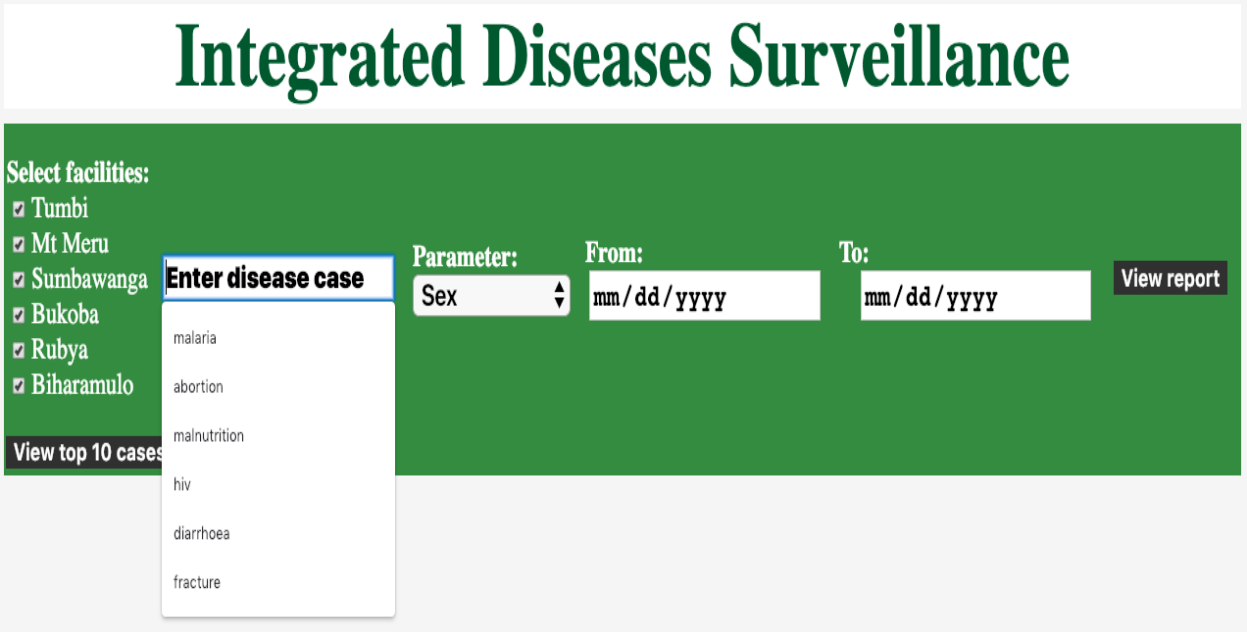


Figure 18: The input field for a disease case.

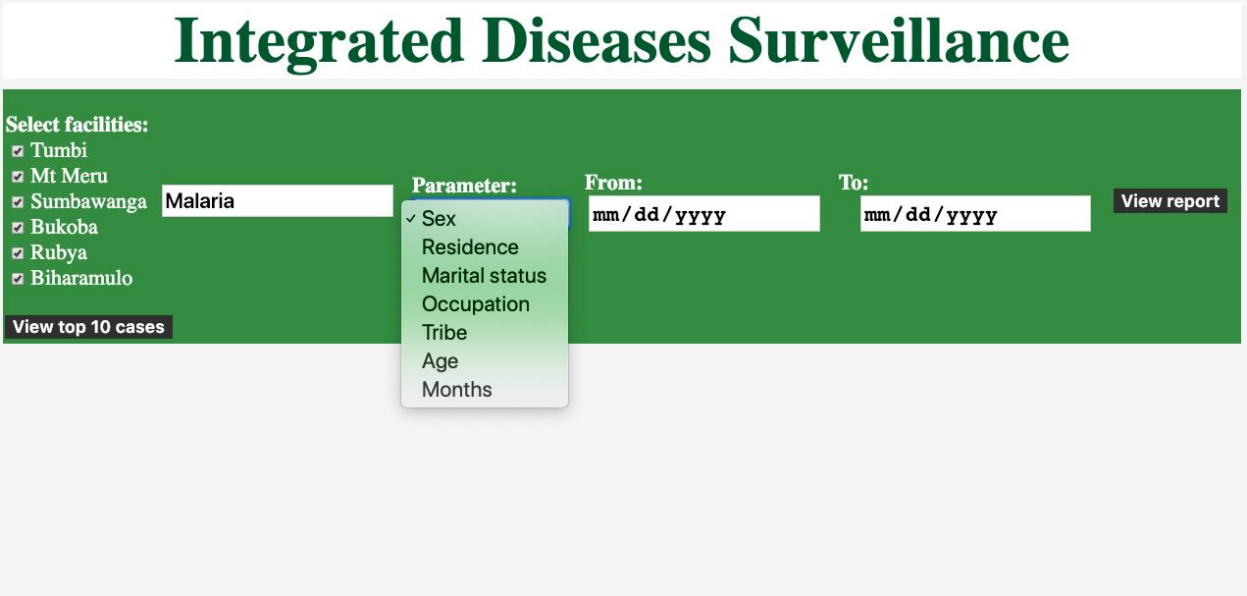


Figure 19: The dropdown menu for attributes for the report.

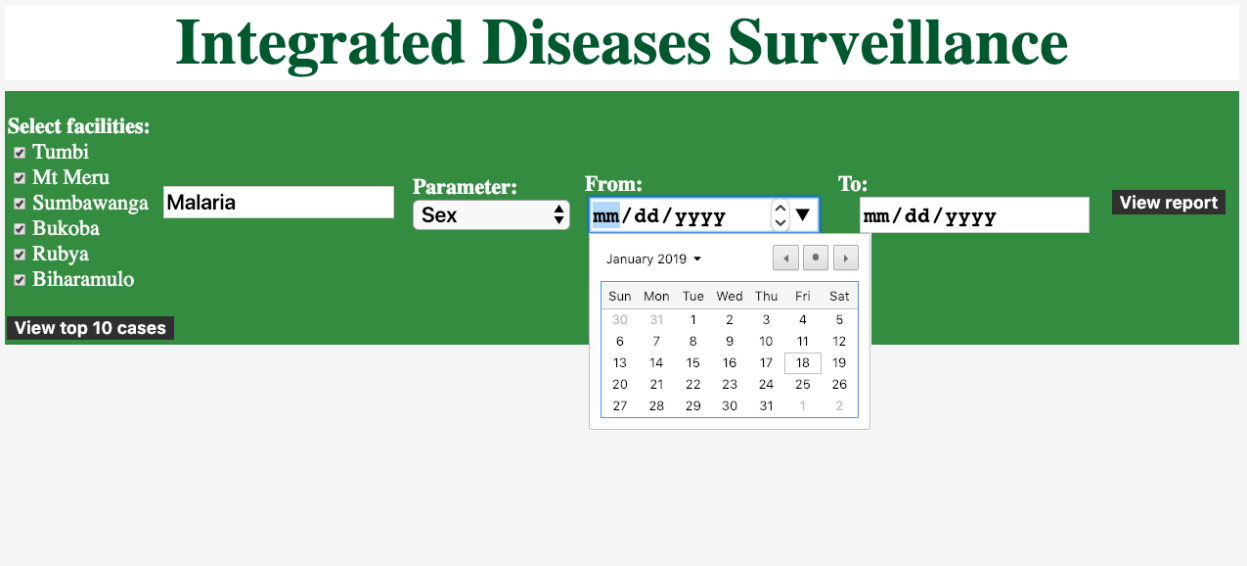


Figure 20: The start-date field to filter the range of data to be analyzed and displayed on the report.

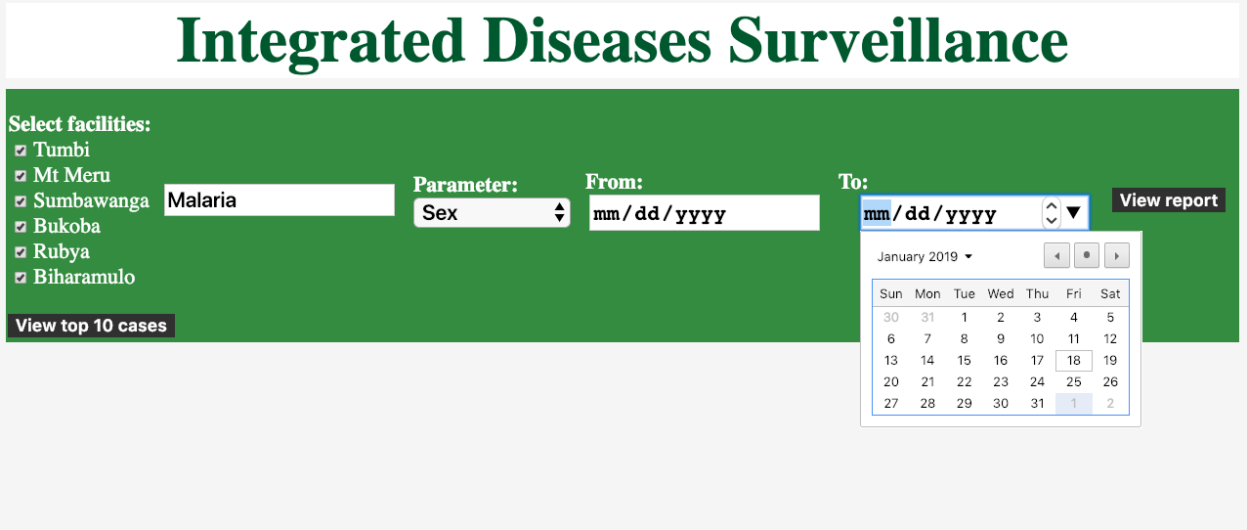


Figure 21: The end-date field to filter the range of data to be analyzed and displayed on the report.

As part of the analysis, an observation on the data in the fact table (fact_diagnosis) revealed a significant number of NULL values in certain attributes, mostly the tribe_id, occupation_id and status_id. This means that tribe, occupation, and marital status data for some patients were not filled during registration process. More than 1.7% of the patients missed the tribe, occupation, and marital status details all together, and 35.9% missed either of the three. This calls for immediate remedy to rectify the faults in patients' registration process and the overall data entry in the system. There are also some faults in the data fed, as a simple test analysis of disease cases versus patients' gender and age revealed 24 cases where gender was male and the disease case was abortion, between 1st September, 2018 and 27th December, 2018; and five abortion cases where age was less than five, as seen in Figs. 22 and 23.

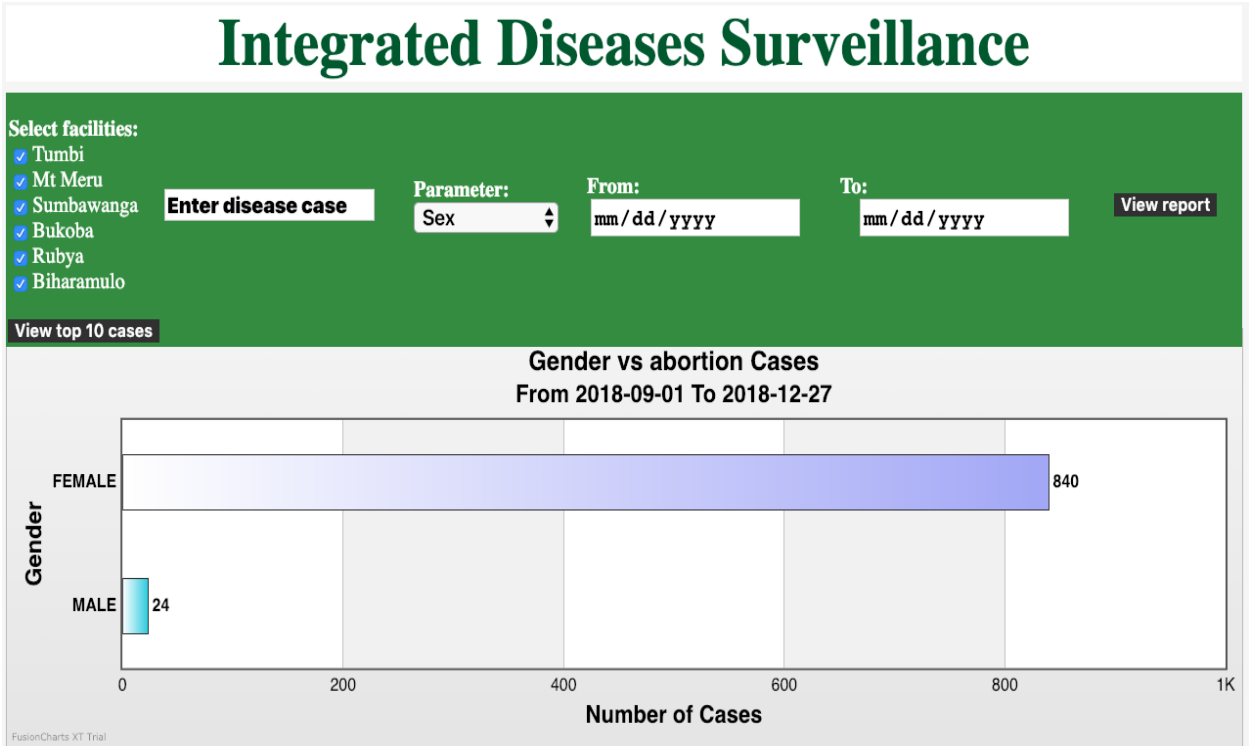


Figure 22: Gender vs abortion cases.

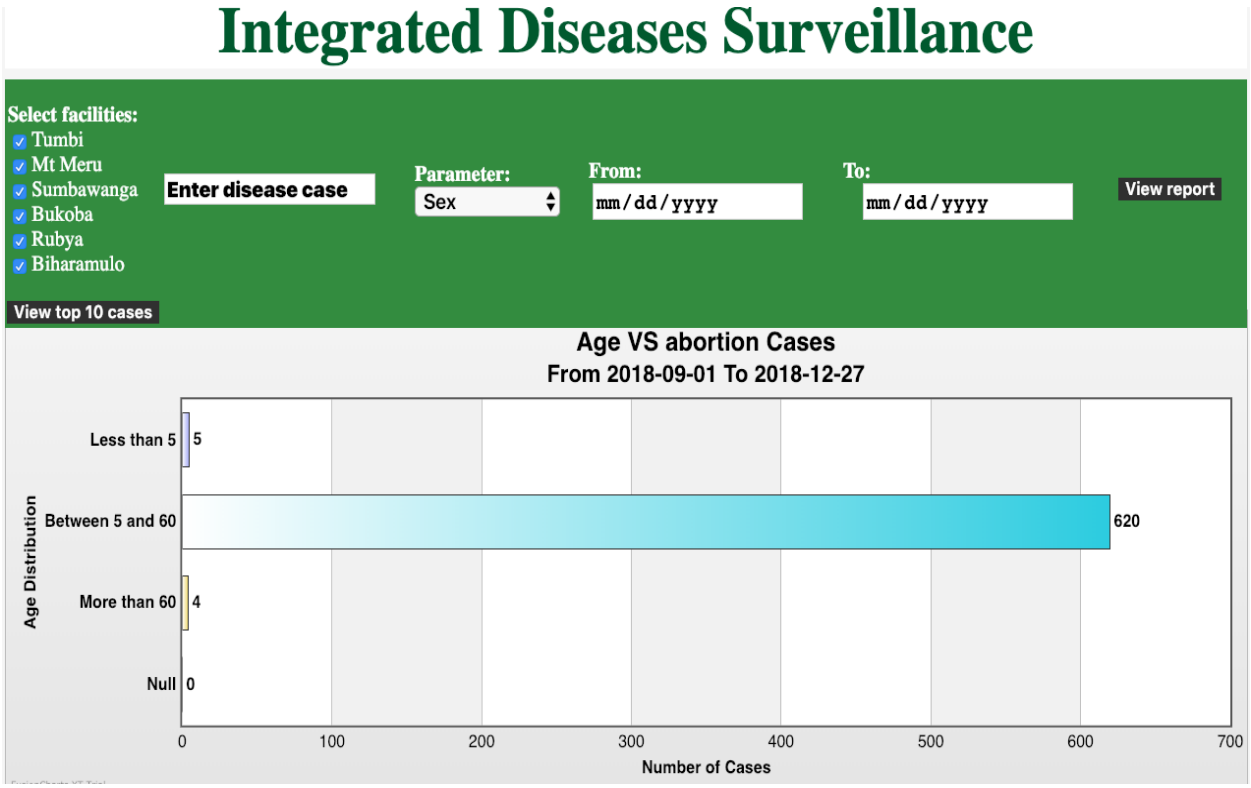


Figure 23: Age vs abortion cases.

Further analysis was done to check the effectiveness of the dataset in generation of reports. Scenarios such as generating the most frequent disease cases (Fig. 24) and simulating a disease case against some of the suggested attributes were tested (Fig. 25 to Fig. 31). Malaria was selected as a testing case due to the fact that it ranked top in the most frequent disease cases chart. However, similar analysis can be done on any disease case. The charts are limited to the top ten records in descending order, based on number of cases. The charts of Fig. 24 through Fig. 31 show several analytics as described in the charts headers.

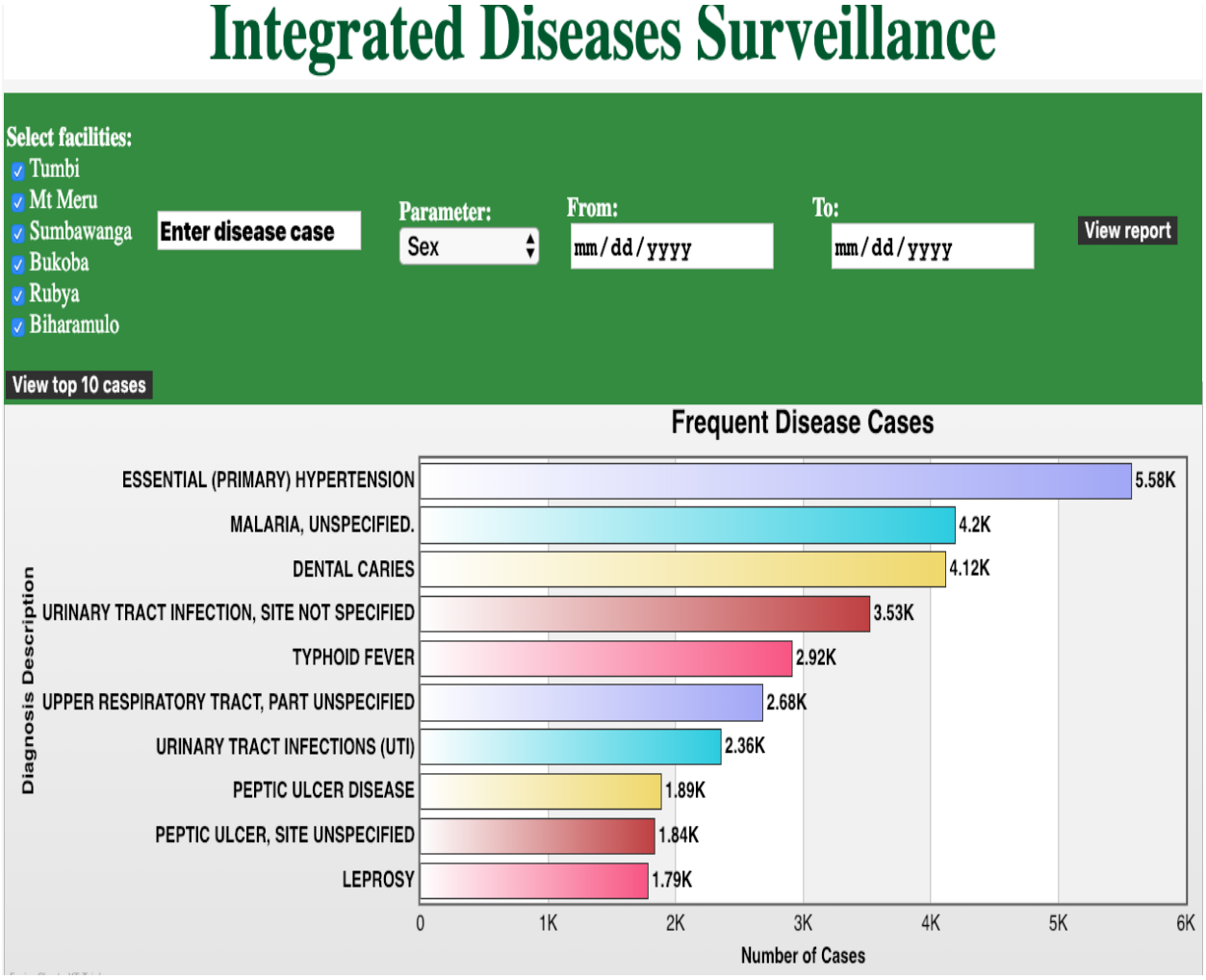


Figure 24: Most frequent disease cases.

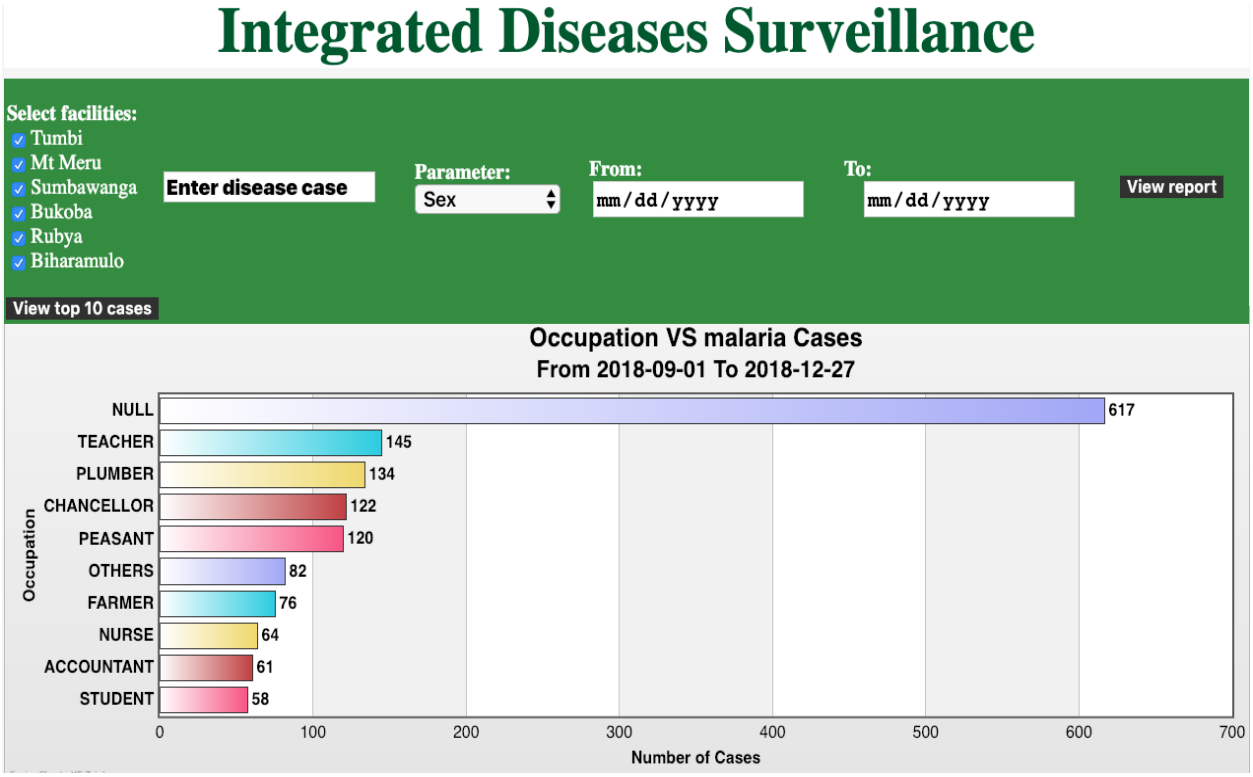


Figure 25: Occupation versus Malaria cases.

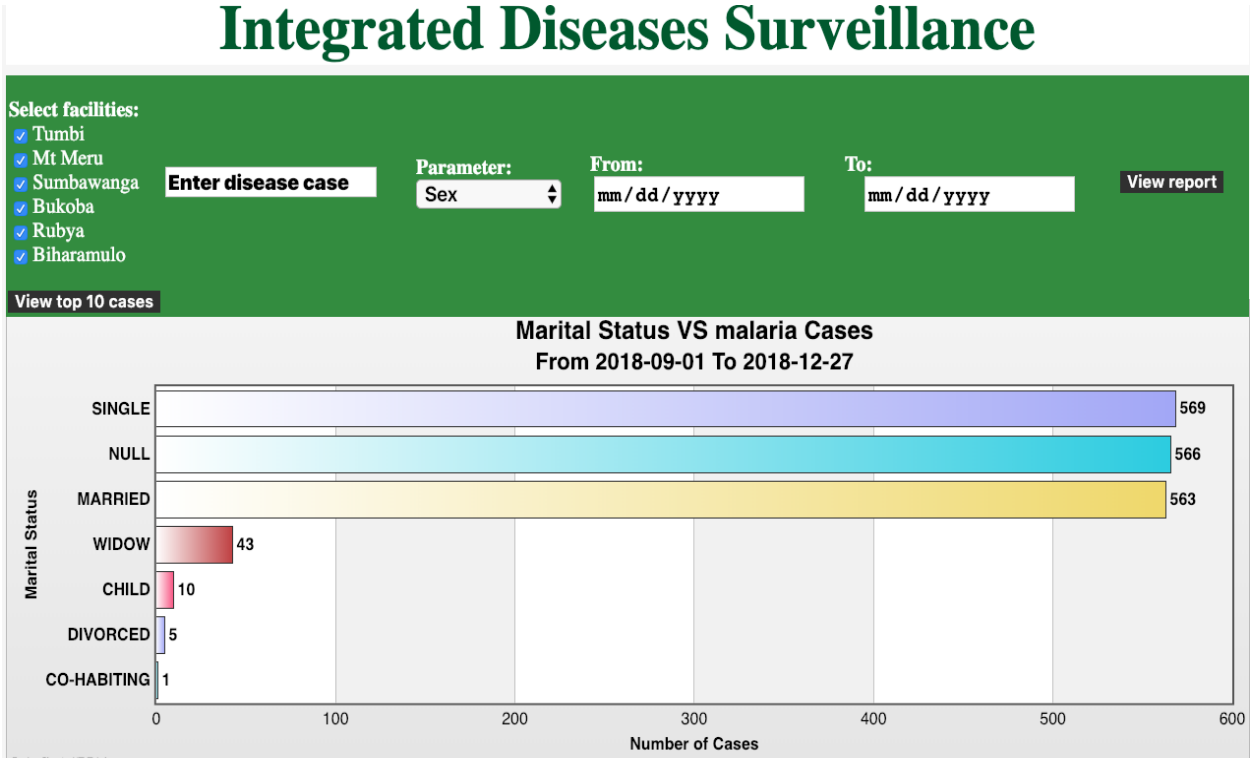


Figure 26: Marital status versus Malaria cases.

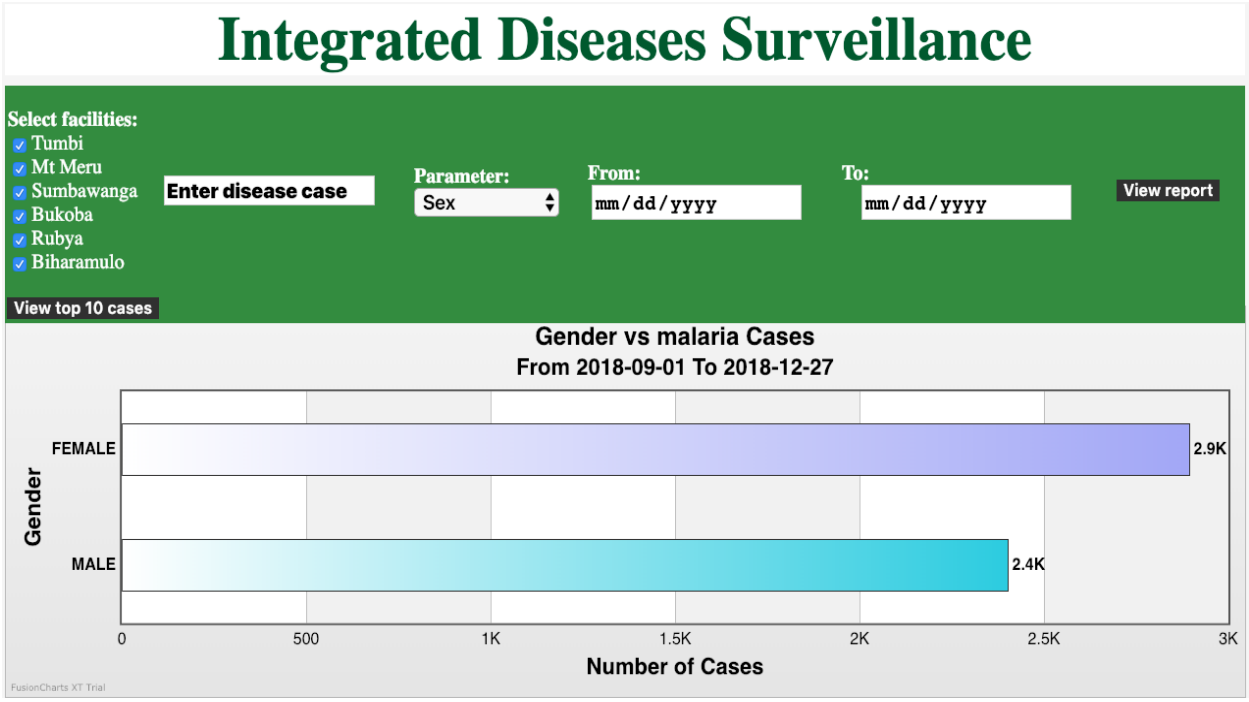


Figure 27: Gender versus Malaria cases.

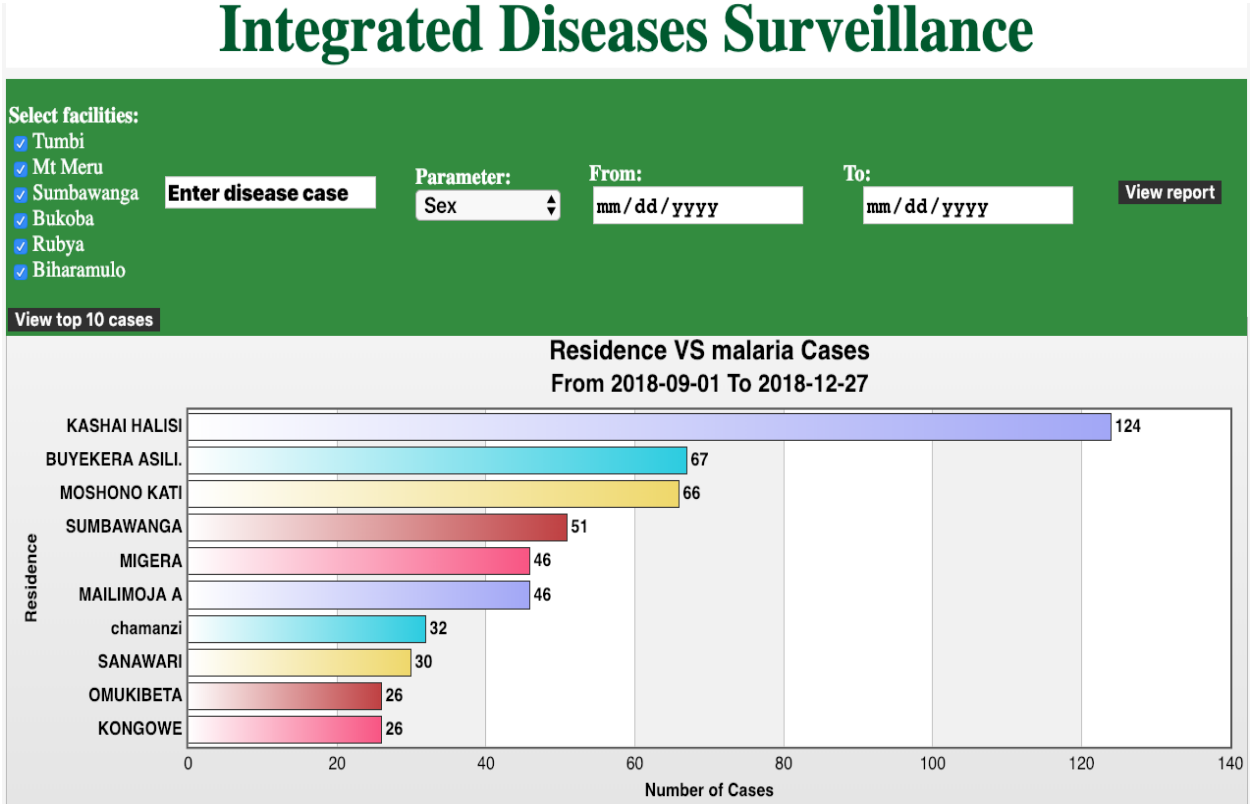


Figure 28: Residence versus Malaria cases.

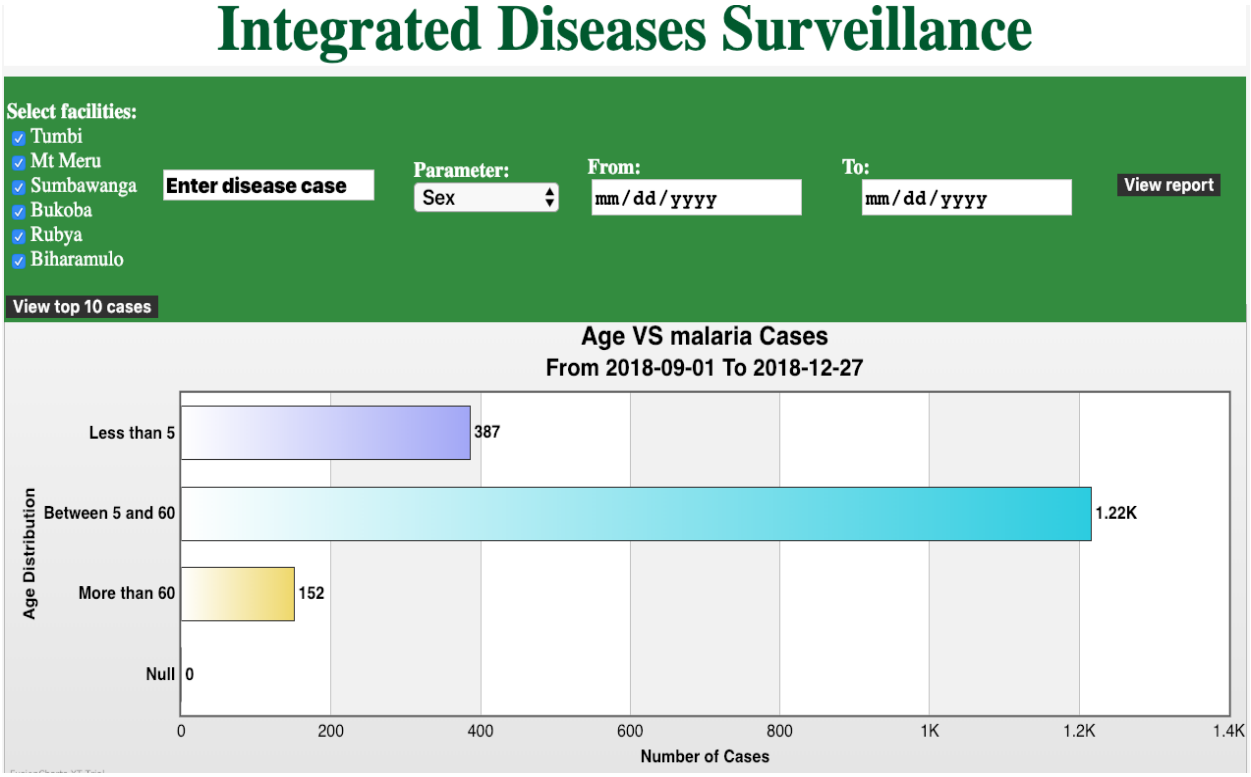


Figure 29: Age versus Malaria cases.

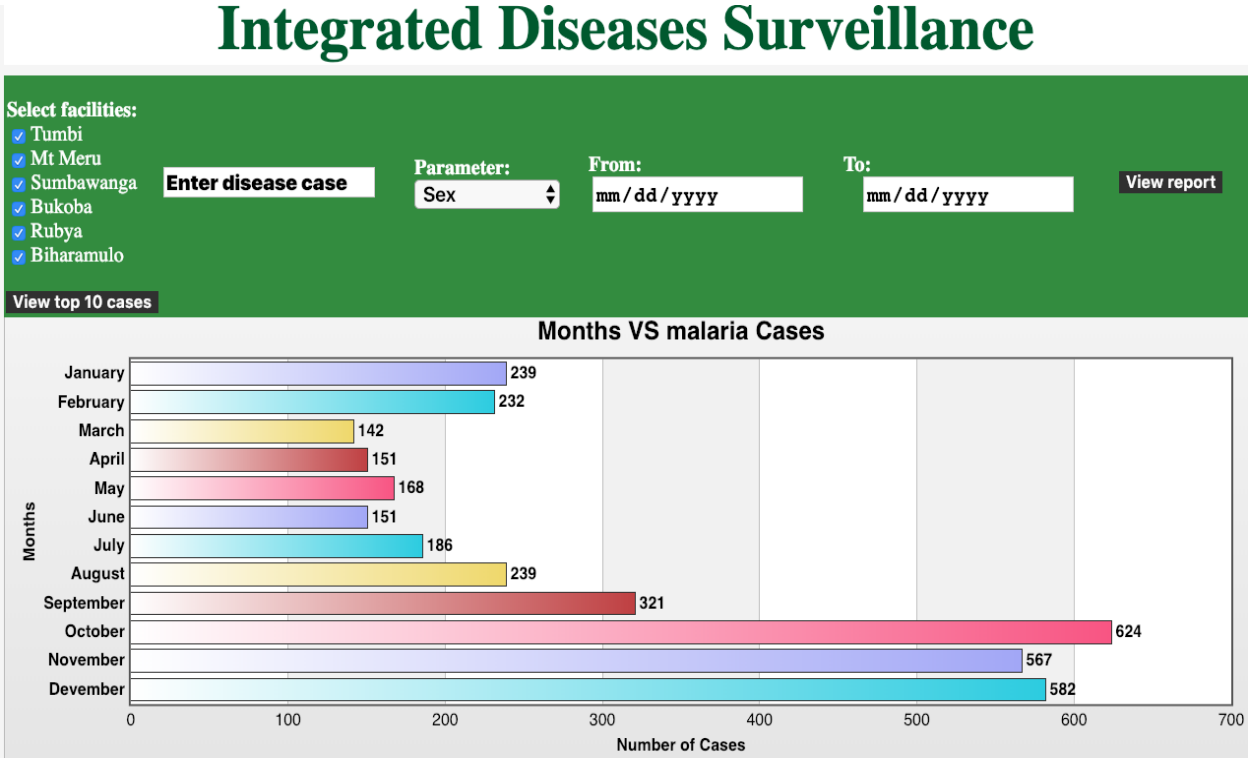


Figure 30: Months versus Malaria cases.

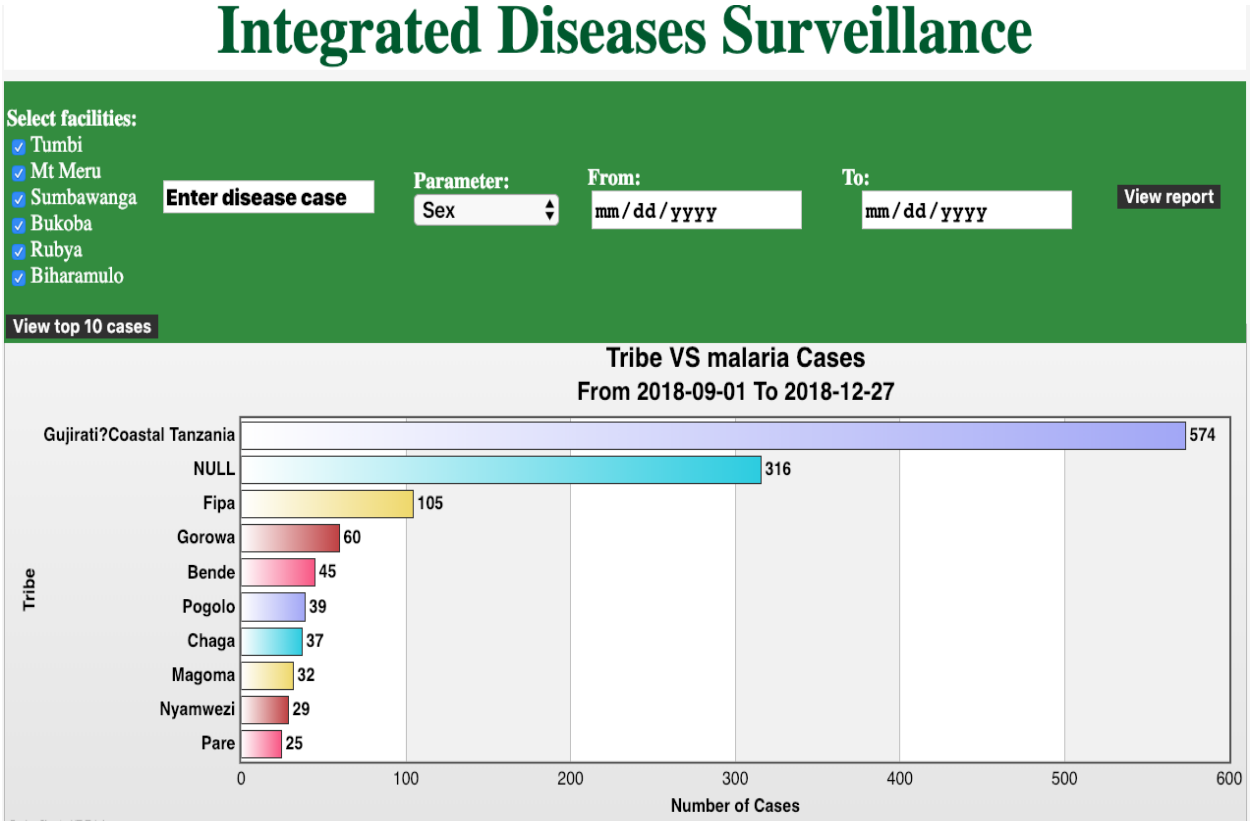


Figure 31: Tribe versus Malaria cases.

The simulated reports are based on Malaria, but any other disease case can be chosen by simply filling its name in the input field for a disease case in the parameters panel. The queries use the LIKE operator, when comparing the user-filled disease cases with the diagnosed diseases whose diagnosis_id are present in the fact table. The LIKE operator searches for string patterns between the user's input and that in the database. It is useful in a sense that a user may be a layman, and simply fill "malaria" as a disease case of interest, not knowing that malaria is further classified in ICD-10 with several variation. Using the LIKE operator followed by a string "%malaria%" (where malaria is the submitted disease case, and % is a wildcard) will search for any diagnosis description having a string pattern matching "malaria". This can be at any position in the diagnosis description string, hence will yield results inclusive of most, if not all variations. On the other hand, using the "=" operator would not yield the best result in our case as this would search for the exact match of the submitted string (Beaulieu, 2009).

The ETL module is also an advancement to the existing setup as it can generate more timely reports. If it is integrated in the GoT-HoMIS, it has the ability to extract the data from the tables

and populate them in the dimension and fact tables timely, depending on the timer set for the scripts to run, either daily at a certain time, or a particular time interval. Once the scripts have populated the fact table, the trends of a disease case can be monitored and visualized as presented above. Day to day monitoring of a disease case is also possible by just selecting the date interval in the user interface, as demonstrated in Fig. 32. This means that during an outbreak, daily progress of an epidemic case can be monitored and reports generated and shared for decision making. The date range is flexible, and hence weekly, monthly, and any other time range reports can be generated as well.

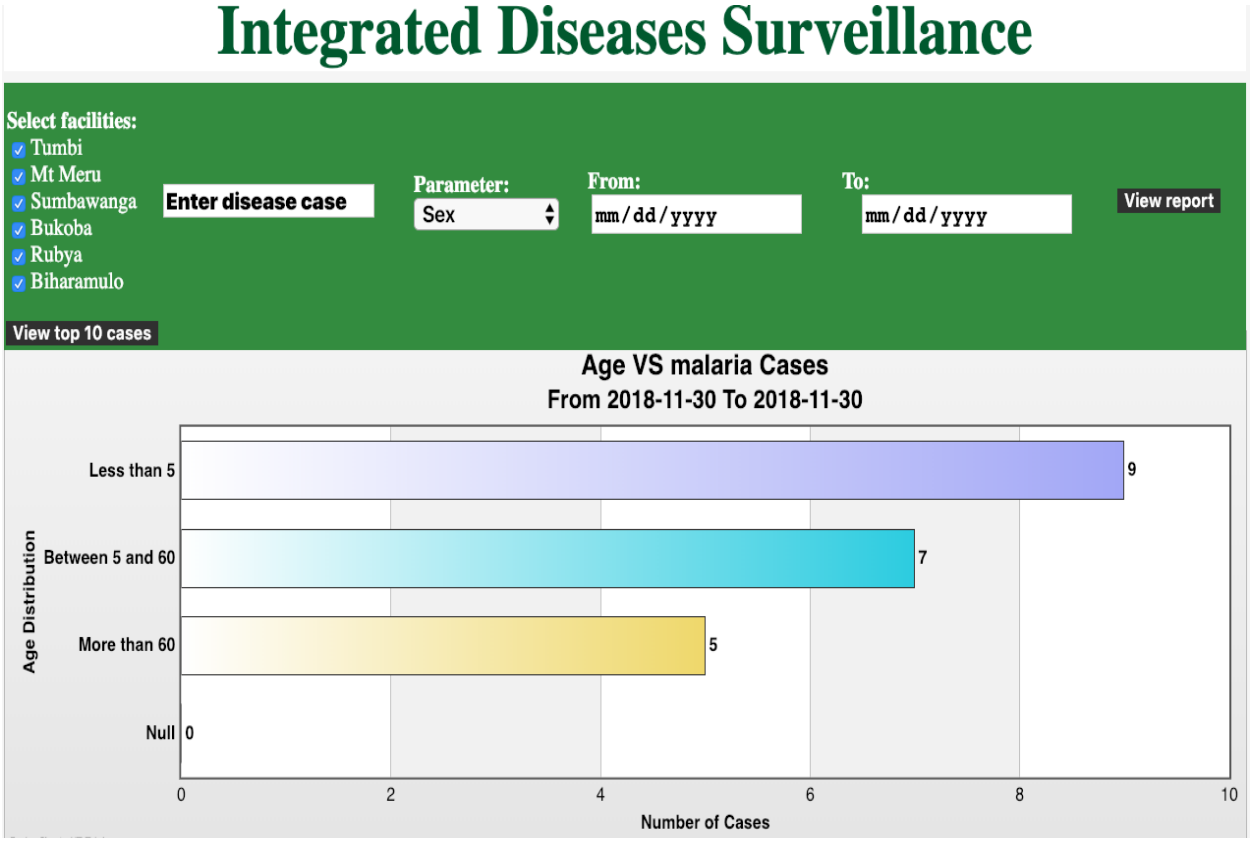


Figure 32: Age analysis on Malaria cases diagnosed in a single day.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This study set out to design a way on which timely clinical data of high quality, that are anonymous, harmonized, and integrated could be obtained so as to enhance epidemic diseases surveillance. A data warehouse was targeted to be the nucleus of the design, where the needed data could be extracted from the source databases to the data warehouse. GoT-HoMIS was selected as a pilot HoMIS to be used in the study. The resulting dataset needed therefore to be sufficient for analysis towards epidemic disease cases.

Three research questions were formed to guide this study towards meeting its objectives. These were: (a) What data attributes stored in the individual databases are essential in facilitating tracking and control of epidemic diseases (first research question)? (b) How best should the data warehouse be designed and implemented to host the extracted and transformed datasets from the multiple sources (second research question)? (c) And, how effective is the resulting dataset in supporting tracking and control of epidemic diseases (third research question)? Answers to the research questions were successfully obtained as the essential attributes to be extracted and used for epidemic diseases surveillance (first research question) were identified to be gender, age, tribe, residence, marital status, occupation, diagnosis, date of onset, vital status, religion, and catchment area population (first specific objective). After the GoT-HoMIS database schema had been thoroughly studied, the data warehouse was designed following the snowflake schema due to the way needed data were mapped in the GoT-HoMIS, and implemented to run on MariaDB database environment as the GoT-HoMIS runs on the same, among other of its merits (second research question; second specific objective). Afterwards, the data warehouse was populated with extracted and transformed data from six (6) health facilities that use the GoT-HoMIS. The resulting dataset was analyzed, and from it different reports based on the user-specified facility(ies), disease cases, attributes recommended for epidemic diseases surveillance, and time stamp can be timely generated to support epidemic diseases surveillance (third research question; third specific objective).

As part of deliverables, an ETL module was developed to extract data from the GoT-HoMIS databases, transform them for harmonization, and load them into the developed data warehouse.

In the GoT-HoMIS nodes, the ETL module can be configured to automatically run at any time interval, which will lead to having more current integrated data for analysis. The analysis can also be configured for data of any interval such as a day, week, month, and so on, as demonstrated.

Further enquiries were made on the current disease cases reporting setup at the MoHCDGEC. While already there exist a number of potential innovations that address various dataset problems, there was still a need for an innovation to extract and analyze the identified parameters against disease cases in a timely manner, with minimal discrepancies, and the analysis be made available to epidemiologists for decision making even before they set feet on the ground. Furthermore, the designed solution considered the possibility of being integrated with meteorological data and community data for enhanced diseases surveillance. A framework for such integration has also been presented and elaborated.

In conclusion, the data warehouse and associated ETL module are therefore strong contributions towards the initiatives for epidemic diseases surveillance in Tanzania, mainly because: (a) they capture the identified parameters in a timely manner; (b) have fewer discrepancies as they fetch them straight from the horse's mouth (the GoT-HoMIS nodes); (c) have the potential to integrate more non-clinical data such as meteorological data (with the date dimension); (d) the data are loaded in non-aggregated manner and hence more advanced analytics can be made that will significantly enhance epidemic diseases surveillance; and (e) the resulting dataset is anonymous and hence can be shared with other interested stakeholders for research purposes. The data integration framework also adds to the on-going initiatives towards strengthening the Tanzania health sector through ICT innovations, by providing a new angle of focus for further research and innovations on timely and comprehensive health data analysis with impact.

5.2 Recommendations

The study recommends the following:

(i) Complete integration of the ETL module in the GoT-HoMIS package and configuring it to run at fixed time intervals so as to generate timely datasets.

(ii) Sharing of the resulting dataset with other researchers and stakeholders for independent analysis, that way more inputs can always be received.

(iii) More hospitals in the United Republic of Tanzania (public and private) to adopt the use of HoMISs for managing hospital operations so that the identified parameters can be timely extracted, and even other data for other analysis and decision making.

(iv) Educating the personnel involved in feeding data in the systems on the essence of quality data and the applications of data for analysis, as this may help improve the quality of data in the systems.

(v) Further research on: (a) the optimization of the algorithms for the developed ETL module; (b) the network infrastructure to enable secure transfer of the extracted data from remote GoT-HoMIS nodes to the data warehouse (central server); (c) integrating the resulting dataset with other electronic HoMISs operated in other facilities, such as Care2x and Jeeva; (d) Mechanisms to capture similar data from health facilities that cannot install electronic HoMISs at the moment, these could also be useful when electricity is off; and (e) integration of meteorological and community data as envisioned on the data integration framework.

REFERENCES

Bartholomew, D. (2012). Mariadb vs. mysql. Retrieved on 13th September, 2017, from http://rozero.webcindario.com/disciplinas/fbmg/abd3/MariaDB_vs_MySQL.pdf.

Beaulieu, A. (2009). *Learning SQL*. Beijing: OReilly.

Cherniack, M., Lawande, S., & Tran, N. (2014). U.S. Patent No. 8,671,091. Washington, DC: U.S. Patent and Trademark Office.

Chute, C. G., Beck, S. A., Fisk, T. B., & Mohr, D. N. (2010). The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. *Journal of the American Medical Informatics Association*, 17(2), 131-135.

Darcy, N. M., Somi, G., Matee, M., Wengaa, D., & Perera, S. (2017). Analysis of Data Dissemination and Use Practices in the Health Sec-tor in Tanzania: Results of desk review and interviews with key stakeholders. *Journal of Health Informatics in Africa*, 4(1).

Edwards, P. (2017). Epidemics: past, present and future – what are the risks? Retrieved on 7th September 2017, from https://www.hannover-rueck.de/1085858/recent-medical-news-epidemics-2017.pdf.

Fernández-Luque L., & Bau T. (2015). Health and social media: perfect storm of information. *Healthcare Informatics Research*, 21(2), 67-73.

Hammergren, T. C., & Simon, A. R. (2009). *Data Warehousing For Dummies* (2nd ed.). Indianapolis, Indiana: Wiley Publishing.

Hays, J. N. (2005). *Epidemics and pandemics: their impacts on human history*. Abc-clio.

Health Metrics Network (HMN). (2008). Framework and standards for country health information systems.

Househ, M. (2016). Communicating Ebola through social media and electronic news media outlets: A cross-sectional study. *Health Informatics Journal*, 22(3), 470-478.

Imhoff, C., Galemmo, N., & Geiger, J. G. (2003). *Mastering data warehouse design: relational and dimensional techniques*. New York: Wiley.

Jovic, A., Brkic, K., & Bogunovic, N. (2014, May). An overview of free software tools for general data mining. In Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on (pp. 1112-1117). IEEE.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, 13(1), 61-72.

Karuri, J., Waiganjo, P., Daniel, O. R. W. A., & Manya, A. (2014). DHIS2: the tool to improve health data demand and use in Kenya. *Journal of Health Informatics in Developing Countries*, 8(1).

Kibaha Education Centre (KEC). (2015). Improving User-fees Collection in Pwani Region [PowerPoint Slides]. Retrieved on 11th September, 2017, from http://ehealth.go.tz/admin/rmo_materials/Day3/Namna%20ya%20kuboresha%20Makusanyo%20-%20 Mganga%20Mkuu,%20PWANI.pdf.

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). Indianapolis, IN: John Wiley & Sons.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management*, 19(2), 65.

Lakshmi, K. K., Gupta, H., & Ranjan, J. (2017, December). USSD—Architecture analysis, security threats, issues and enhancements. In Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS), 2017 International Conference on (pp. 798-802). IEEE.

Lee, H. F., Fei, J., Chan, C. Y., Pei, Q., Jia, X., & Yue, R. P. (2017). Climate change and epidemics in Chinese history: A multi-scalar analysis. *Social Science & Medicine*, *174*, 53-63.

Mandara, M., Kapinga, A., Malisa, G., Minja, G., Kajeguka, A., & Junga, L. (2005). HMIS Assessment in Mtwara Region and Proposal for Strengthening the MTUHA System.

Marchant, T., Schellenberg, J., Peterson, S., Manzi, F., Waiswa, P., Hanson, C., ... & Rowe, A. K. (2014). The use of continuous surveys to generate and continuously report high

quality timely maternal and newborn health data at the district level in Tanzania and Uganda. *Implementation Science*, *9*(1), 112.

Merrill, R. M. (2015). *Introduction to epidemiology*. Jones & Bartlett Publishers.

Ministry of Health and Social Welfare (MoHSW). (2013). Tanzania National eHealth Strategy June, 2013 – July, 2018. Retrieved on 11th September, 2017, from http://www.tzdpg.or.tz/fileadmin/documents/dpg_internal/dpg_working_groups_clusters/cluster_2/health/Key_Sector_Documents/Tanzania_Key_Health_Documents/Tz_eHealth_Strategy_Final.pdf.

Ministry of Health and Social Welfare (MoHSW). (2009). Proposal To Strengthen Health Information System (HIS). Retrieved on 12th September, 2017, from http://www.tzdpg.or.tz/fileadmin/documents/dpg_internal/dpg_working_groups_clusters/cluster_2/health/Key_Sector_Documents/Monitoring___Evaluation/Proposal_to_Strengthen_Health_Information_System.pdf.

Ministry of Health, Community Development, Gender, Elderly and Children (MoHCDGEC). (2018). Tanzania e-IDSR System [PowerPoint Slides].

Ministry of Health, Community Development, Gender, Elderly, and Children (MoHCDGEC). (2016). National Guidelines for Health Data Quality Assessment. Retrieved on 2nd January, 2019, from https://www.medbox.org/national-guidelines-for-health-data-quality-assessment/download.pdf.

Morais, A., Peixoto, H., Coimbra, C., Abelha, A., & Machado, J. (2017). Predicting the need of Neonatal Resuscitation using Data Mining. *Procedia Computer Science*, 113, 571-576.

Munirathinam, S., & Ramadoss, B. (2016). Predictive models for equipment fault detection in the semiconductor manufacturing process. *International Journal of Engineering and Technology*, 8(4), 273-285.

Naderifar, M., Goli, H., & Ghaljaie, F. (2017). Snowball sampling: A purposeful method of sampling in qualitative research. *Journal of Strides in Development of Medical Education*, 14(3).

Nikfarjam, A., Sarker, A., O'connor, K., Ginn, R., & Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with

word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3), 671-681.

Nyasulu, P., Kasubi, M., Boniface, R., & Murray, J. (2014). Understanding Laboratory Methods and Their Impact on Antimicrobial Resistance Surveillance, at Muhimbili National Hospital, Dar es Salaam, Tanzania. *Advances in Microbiology*, 4(01), 33.

Paul, M. J., Sarker, A., Brownstein, J. S., Nikfarjam, A., Scotch, M., Smith, K. L., & Gonzalez, G. (2016). Social media mining for public health monitoring and surveillance. In Biocomputing 2016: Proceedings of the Pacific Symposium(pp. 468-479).

Polit, D. F., & Beck, C. T. (2010). *Essentials of nursing research: appraising evidence for nursing practice* (7th ed., pp.309; 319). Philadelphia, PA: Wolters Kluwer.

President's Office – Regional Administration and Local Government (PO-RALG). (2017). Vituo 170 vya Kutolea Huduma za Afya Vyafungiwa Mfumo wa GoTHoMIS. Retrieved on 28th September, 2017, from http://fullshangweblog.com/home/2017/09/21/vituo-170-vya-kutolea-huduma-za-afya-vyafungiwa-mfumo-wa-gothomis/.

Ross, T. R., Ng, D., Brown, J. S., Pardee, R., Hornbrook, M. C., Hart, G., & Steiner, J. F. (2014). The HMO Research Network Virtual Data Warehouse: a public data model to support collaboration. *eGEMS*, 2(1).

Rumisha, S. F., Mboera, L. E., Senkoro, K. P., Gueye, D., & Mmbuji, P. K. (2007). Monitoring and evaluation of integrated disease surveillance and response in selected districts in Tanzania. *Tanzania Journal of Health Research*, 9(1), 1-11.

Rutatola, E. P., Yonah, Z. O., Nyambo, D. G., Mchau, G. J., & Musabila, A. K. (2018). A Framework for Timely and More Informative Epidemic Diseases Surveillance: The Case of Tanzania. *Journal of Health Informatics in Developing Countries*, 12(2).

Sæbø, J. I., Kossi, E. K., Titlestad, O. H., Tohouri, R. R., & Braa, J. (2011). Comparing strategies to integrate health information systems following a data warehouse approach in four countries. *Information Technology for Development*, 17(1), 42-60.

Sen, A., & Sinha, A. P. (2005). A comparison of data warehousing methodologies. *Communications of the ACM*, 48(3), 79-84.

Shen, J. C., Lei, L. U. O., Li, L. I., Jing, Q. L., OU, C. Q., Yang, Z. C., & Chen, X. G. (2015). The impacts of mosquito density and meteorological factors on dengue fever epidemics in Guangzhou, China, 2006–2014: a time-series analysis. *Biomedical and Environmental Sciences*, *28*(5), 321-329.

Supaartagorn, C. (2016). A Framework for Web-based Data Visualization using Google charts based on MVC pattern. King Mongkut's University of Technology North Bangkok. *International Journal of Applied Science and Technology*, 9(4).

Szklo, M., & Nieto, J. (2014). *Epidemiology*. Jones & Bartlett Publishers.

Tao, X. Y., Tang, Q., Rayner, S., Guo, Z. Y., Li, H., Lang, S. L., ... & Song, M. (2013). Molecular phylodynamic analysis indicates lineage displacement occurred in Chinese rabies epidemics between 1949 to 2010. *PLoS Neglected Tropical Diseases*, *7*(7), e2294.

Tommy. (2017). Date Dimension File. Retrieved July 26, 2018, from https://support.sisense.com /hc/en-us/articles/230644208-Date-Dimension-File.

Vaisman, A., & Zimányi, E. (2014). *Data Warehouse Systems: Design and Implementation*. Springer.

Van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., ... & Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. *BMC Public Health*, 14(1), 1144.

Walia, E. S., & Gill, E. S. K. (2014). A framework for web based student record management system using PHP. *International Journal of Computer Science and Mobile Computing*, 3(8), 24-33.

Wambura, W., Machuve, D., & Nykänen, P. (2017). Development of Discharge Letter Module onto a Hospital Information System. *Journal of Health Informatics in Developing Countries*, 11(2).

World Health Organization (WHO). (2016). WHO Tanzania Annual Report 2016. Retrieved on 8th May, 2018, from http://www.afro.who.int/publications/who-country-office-annual-report-2016.

World Health Organization. (2016). African Health Observatory. Retrieved on 12th September, 2017, from http://www.aho.afro.who.int/profiles_information/index.php/Tanzania:Analytical_summary_-_Health_information,_research,_evidence_and_knowledge.

APPENDICES

Appendix 1: Interview Guide: Doctors

Development in the field of machine learning has led to the availability of enhanced sophisticated tools and algorithms for data science and visualization that, when applied on clinical data, makes it relatively easier to analyze and discover patterns that show the nature and characteristics of an epidemic and hence enhancing epidemiology. Despite such development, availability of localized datasets for sharing amongst stakeholders and applying such enhanced algorithms on the datasets is still a challenge. This study therefore intends to come up with a software module for extracting and transforming patients' clinical and demographic data from the Government of Tanzania – Hospital Management Information System (GoT-HoMIS) and loading them into a dedicated data warehouse to enable the application of machine learning algorithms to identify patterns embedded in the epidemic cases.

Through this 30 minutes interview, the researchers aim to identify and understand the essential data attributes (both clinical and demographic) that are to be extracted from the individual GoT-HoMIS databases to enhance analysis on the epidemics. It is our hope that you will be pleased to be a part of this journey. We would like to confirm that data provided in this study will be used for academic purposes only and according to ethical standards observed by all research institutions. Please be assured that highest degree of confidentiality and anonymity of respondents will be observed.

SECTION A: About yourself

1. Name:…… (optional)
2. Age group (years):

20 - 30	31 - 35	36 - 40	41 - 45	More than 45

3. Sex: Male/Female
4. Position/title: …….
5. Institution/Organisation: …….
6. Working experience (years): ……...
7. Years worked with GoT-HoMIS:……
8. Number of Institutions ever worked for: ……
9. Highest Education level: ……

SECTION B: The interview

1. Which of the patients' particulars recorded in the GoT-HoMIS (attached herewith) can be used for epidemics surveillance?
2. Which of the particulars discussed above can you term as identity revealing?
3. Can the particulars be classified as primary and secondary in terms of necessity?
4. Any suggestions for modification on the information collected (collected particulars, format)?
5. Any suggestions for modification on the epidemic cases analytics and feedback process in the GoT-HoMIS?
6. Is there any other needed information that is not currently captured in the GoT-HoMIS?
7. Do you see a prospect of having a central epidemics surveillance system with GoT-HoMIS as one of the data sources? Any suggestions on how to get there?
8. Any reflections on the general operations of the GoT-HoMIS?
9. Any suggestion, question, advice that can help the research?

Appendix 2: Interview Guide: Epidemiologists

Development in the field of machine learning has led to the availability of enhanced sophisticated tools and algorithms for data science and visualization that, when applied on clinical data, makes it relatively easier to analyze and discover patterns that show the nature and characteristics of an epidemic and hence enhancing epidemiology. Despite such development, availability of localized datasets for sharing amongst stakeholders and applying such enhanced algorithms on the datasets is still a challenge. This study therefore intends to come up with a software module for extracting and transforming patients' clinical and demographic data from the Government of Tanzania – Hospital Management Information System (GoT-HoMIS) and loading them into a dedicated data warehouse to enable the application of machine learning algorithms to identify patterns embedded in the epidemic cases.

Through this 30 minutes interview, the researchers aim to identify and understand the essential data attributes (both clinical and demographic) that are to be extracted from the individual GoT-HoMIS databases to enhance analysis on the epidemics. It is our hope that you will be pleased to be a part of this journey. We would like to confirm that data provided in this study will be used for academic purposes only and according to ethical standards observed by all research institutions. Please be assured that highest degree of confidentiality and anonymity of respondents will be observed.

SECTION A: About yourself

10. Name:…… (optional)
11. Age group (years):

20 - 30	31 - 35	36 - 40	41 - 45	More than 45

12. Sex: Male/Female
13. Position/title: …….
14. Institution/Organisation: …….
15. Working experience (years): ……...
16. Number of Institutions ever worked for: ……
17. Highest Education level: ……

SECTION B: The interview

10. How do you get notifications and information about disease cases? From where/whom? The process?
11. Are you familiar with the Government of Tanzania Hospital Management Information System (GoT-HoMIS)?
12. Which of the patients' particulars recorded in the GoT-HoMIS (attached herewith) can be used for epidemics surveillance?
13. Can the particulars above be specified as primary or secondary in terms of necessity?
14. Which of the above patients' particulars can you term as identity revealing?
15. Is there any additional information needed but not included in the attached list?
16. What information (reports) would be helpful in epidemics surveillance from disease cases datasets?
17. How do you currently analyze different disease cases from multiple health facilities?
18. What are the reports produced from the analysis process? What are the issues included in the reports? Are there graphical visualizations of the reported issues? How are they shared?
19. Any suggestions for improvement in the current analysis process, systems, and/or reports?
20. Any suggestions for modification on the information collected (collected particulars, format)?
21. Do you see a prospect of having a central epidemics surveillance system with GoT-HoMIS as one of the data sources? Any suggestions on how to get there?

22. Are there any systems that can be integrated with the GoT-HoMIS in order to enhance the real-time surveillance and analysis of disease cases from multiple health facilities?
23. Any suggestion, question, advice that can help the research?

Appendix 3: List of patient's particulars collected in the GoT-HoMIS registration form

1. First name.
2. Middle name.
3. Surname.
4. Gender.
5. Date of birth.
6. Age.
7. Mobile number.
8. Tribe.
9. Residence.
10. Marital status.
11. Occupation.
12. Country.
13. Next of kin (name, residence, relationship, phone number).

Appendix 4: GoT-HoMIS Access Permit

THE UNITED REPUBLIC OF TANZANIA
PRESIDENTS OFFICE

Telegrams: "TAMISEMI"DODOMA
Telephone: (026) 2322848,
2321607,2322853, 2322420
Fax No: (026) 2322116, 2322146,
2321013
E-mail: ps@porag.go.tz
In reply please quote:
Ref. No. AB.81/228/01

Regional Administration
And Local Government,
Mkapa House,
Hospital Street,
P. O. Box 1923,
41185 DODOMA.

07th December, 2018

Edgar Rutatola

Re: **REQUEST TO ACCESS GoTHOMIS DATA FOR RESEARCH AND DEVELOPMENT OF ADDED MODULE FOR REPORTS ON EPIDEMICS**

The above captioned subject is referred to.

Reference is made to your letter dated 29th August, 2018 that bears the above title. I'm indebted to inform you that the permission to access GoTHOMIS platform on specifically epidemics aggregate data to enable you to fulfill your academic (research) requirements is granted. However, you are required to communicate with the ICT department at the PORALG for more analysis over what shall be accessed as system access protocols will be observed.

Yours,

Dr. Ntuli A. Kapologwe
For: PERMANENT SECRETARY

CC: Shubi kaijage, PhD Ag. Dean

RESEARCH OUTPUT

i. Journal Papers

One research paper titled "**A Framework for Timely and More Informative Epidemic Diseases Surveillance: The Case of Tanzania**" has been published in the Journal of Health Informatics in Developing Countries:

Rutatola, E. P., Yonah, Z. O., Nyambo, D. G., Mchau, G. J., & Musabila, A. K. (2018). A Framework for Timely and More Informative Epidemic Diseases Surveillance: The Case of Tanzania. ***Journal of Health Informatics in Developing Countries*, *12*(2)**.

The paper is also accessible through the link:
http://www.jhidc.org/index.php/jhidc/article/view/185.

ii. Manuscripts

A manuscript titled "**Design and Development of a Data Warehouse for Integrated Monitoring of Disease Cases: The case of Tanzania"** has been prepared and is ready to be submitted for peer review.

Journal of Health Informatics in Developing Countries

http://www.jhidc.org/
Vol. 12 No. 2, 2018

Submitted: July 29th, 2017 Accepted: October10th, 2018

A Framework for Timely and More Informative Epidemic Diseases Surveillance: The Case of Tanzania

Edger P. Rutatola[1], Zaipuna O. Yonah[1], Devotha G. Nyambo[1], Geofrey J. Mchau[1], Albogast K. Musabila[2]

[1]The Nelson Mandela African Institution of Science and Technology, P.O Box 447 Arusha, Tanzania
[2]Mzumbe University, P.O Box 1 Mzumbe, Morogoro, Tanzania

Abstract:

Background: A number of health facilities in the United Republic of Tanzania use different Hospital Management Information Systems (HoMISs) for capturing and managing clinical and administrative information for report generation. Despite the potentials of the data in the systems for use in epidemic diseases surveillance, timely extraction of the data for integrated data mining and analysis to produce more informative reports is still a challenge. This paper identifies the candidate data attributes for epidemic diseases surveillance to be extracted and analyzed from the Government of Tanzania Hospital Management Information System (GoT-HoMIS). It also examines the current reporting setup for epidemic diseases surveillance in Tanzania from the health facilities to the district, regional, and national levels.

Methods: The study was conducted at the Ministry of Health, Community Development, Gender, Elderly, and Children (MoHCDGEC), Tumbi Designated Regional Referral Hospital (TDRRH), Muhimbili University of Health and Allied Sciences (MUHAS), and Mzumbe Health Centre, all in the United Republic of Tanzania. A total of 10 key informants (medical doctors, epidemiologists, and focal persons for various health information systems in Tanzania) were interviewed to obtain primary data. Data entry process in the GoT-HoMIS was also observed. Documents were reviewed to broaden understanding on several aspects.

Results: All the respondents (100%) suggested patients' gender, age, and residence as suitable attributes for epidemic diseases surveillance. Other suggested attributes were occupation (85.71%), diagnosis (57.14%), catchment area population (57.14%), vital status (57.14%), date of onset (57.14%), tribe (42.86%), marital status (42.86%), and religion (14.29%). Timeliness, insufficient immediate particulars on an epidemic-prone case(s), aggregated data limiting extensive analytics, missing community data and ways to analyze rumors, and poor data quality were also identified as challenges in the current reporting setup.

Conclusion: A framework is proposed to guide researchers in integrating data from health facilities with those from social media and other sources for enhanced epidemic disease surveillance. Data entrants in the systems should also be informed on the essence and applications of data they feed, as quality data are the roots of quality reports.

Keywords: District Health Information System; Integrated Disease Surveillance and Response; Epidemic Diseases Tanzania; Data Integration; Health Datasets; Health Management Information System; Data Mining, Data Warehouse.

[1] Edger P. RUTATOLA; The Nelson Mandela African Institution of Science and Technology, P.O BOX 447 Arusha, Tanzania; Tel: +255 658 526 256; Email: rutatolae@nm-aist.ac.tz; eprutatola@mzumbe.ac.tz; edgerrutatola@gmail.com

1. Introduction

Increase in number of local software developers and availability of open source software has led to improved development and management of various Hospital Management Information Systems (HoMISs) in developing countries [1, 2]. In the United Republic of Tanzania (URT), a number of Hospitals (public and private) operate different Information Systems for storage and manipulation of clinical and administrative information [2, 3, 4]. Adoption of the Information systems is made possible due to the support and emphasis from the Government on development and operationalization of such systems. Notable advantages have been gained from the systems, including but not limited to generation of reports that help in hospital administration as well as monitoring clinical operations. Moreover, successive systems such as the District Health Information System (DHIS2) have been installed to generate more comprehensive and integrated reports on different aspects including epidemics surveillance from districts to national level [5]. Ideally, these systems are expected to generate quality and vast datasets for diversified analytical purposes, as quality data is a success factor for generation of useful reports [6]. Good quality data can be used to find demographic and clinical patterns in disease cases from multiple hospitals and provide more insights to epidemiologists as well as detailed and reliable statistical based reports such as the Tanzania Malaria Indicator Survey, 2017 [7, 8].

One among the Hospital Management Information Systems installed in Tanzanian health facilities is the Government of Tanzania – Hospital Management Information System (GoT-HoMIS). It is a system developed by local experts and currently installed in more than one hundred and seventy (170) health facilities [2]. While such a wide adoption of the system could be advantageous for data mining and analytics, the GoT-HoMIS nodes are currently not centralized and hence analysis of collected data is only limited to the host facility. At present, aggregated clinical data from all health facilities in Tanzania are integrated and stored in the DHIS2 on a monthly basis, where analytics are performed, and reports are generated [9]. From the aggregated data, only limited analytics and patterns can be observed. Epidemics surveillance needs to be timely and comprehensive so that actions towards an epidemic are effective [10].

This paper explores the potentials of the patients' clinical and demographic data collected by the GoT-HoMIS for epidemics surveillance. It also examines existing systems and setup for epidemic diseases surveillance in Tanzania, from where a case is confirmed at the health facility to when reports are shared at the district, regional, and national epidemiological teams. Furthermore, a framework for integrating health datasets generated from different health facilities, community data gathered from rumors, and other environmental data such as meteorological data is proposed. With comprehensive integrated dataset from the diverse sources machine learning algorithms can be applied for enhanced analytics to aid epidemics surveillance and informed decision making. The proposed framework is expected to guide researchers on enabling the application of data mining tools on the resulting datasets to

promote generation and dissemination of more comprehensive and timely reports for epidemics surveillance and control.

2. Methods

2.1 Study population

The study population comprised of medical doctors, epidemiologists, health information systems focal persons, and health information systems administrators. These were either based at the Ministry of Health, Community Development, Gender, Elderly, and Children (MoHCDGEC), or among three health facilities located in Morogoro, Dar es Salaam, and Pwani Regions in the United Republic of Tanzania. The chosen health facilities were Tumbi Designated Regional Referral Hospital (TDRRH) (Pwani Region - Tanzania), Mzumbe Health Centre (Morogoro Region - Tanzania), and the Muhimbili University of Health and Allied Sciences (MUHAS) (Dar es Salaam Region - Tanzania). These were purposively selected based on their roles and position in the Tanzanian health sector (MoHCDGEC); experience in operating HoMISs, specifically the GoT-HoMIS (TDRRH and Mzumbe Health Centre); and having a nominated HoMIS pioneer (MUHAS).

A total of ten (10) representative respondents from the mentioned facilities were involved in the study. These were obtained through snowball sampling, based on their key roles and knowledge on epidemiology, and experience in working with the GoT-HoMIS or other Health Management Information systems and innovations in the country. Seven (7) of the ten respondents were medical doctors and/or epidemiologists, who were further divided into two categories. The first category is of those who are stationed in hospitals. This category had four (4) medical doctors; one being stationed at Mzumbe Health Centre, two at TDRRH, and one from MUHAS. These were chosen following their experience in working with HoMISs, particularly the GoT-HoMIS. The second category (Team) comprised of doctors and epidemiologists stationed at the MoHCDGEC in Dar es Salaam, Tanzania. This category had three (3) doctors/epidemiologists; one being the national IDSR focal person, another one being the Public Health Emergency Operations Centre (PHEOC) manager, and lastly a full-time epidemiologist. This category was selected based on their key knowledge on epidemiology and their roles in the current systems and innovations in support of the same.

The remaining three (3) respondents were not of the medicine or epidemiological background. One was a statistician at the MoHCDGEC and a national DHIS2 focal person. The other two were systems administrators having more than four years of experience working with the GoT-HoMIS.

2.2 Data Collection

Semi-structured interviews were conducted to obtain primary data. The interviews intended to identify key attributes to be extracted from the GoT-HoMIS nodes in the health facilities that can help in facilitating timely monitoring and analysis of epidemic cases. The other part of the interviews, specifically dedicated to respondents from the MoHCDGEC, aimed to derive information on the current setup for reporting on disease cases, especially epidemics, from when the diagnosis is made at the health facility to the district, regional, and national responsible teams.

Observation method was also employed on the patients' registration processes at the Tumbi Designated Regional Referral Hospital. Firsthand observation was carried out on how patients were registered over time. The objective was to examine the quality of the data fed into the GoT-HoMIS in terms of correctness and completeness.

Moreover, documents were also collected and reviewed to obtain additional knowledge. Review was done on literature and various documents focusing on health information systems and the setup for reporting and monitoring in the Tanzania's health sector. Some of the documents were obtained online, whereas some were provided by focal persons at the MoHCDGEC.

3. Findings

3.1 Respondents' Profiles

Seven (7) of the ten respondents (medical doctors and/or epidemiologists) had academic qualifications and working experiences as presented in the charts of Figs. 1 and 2:

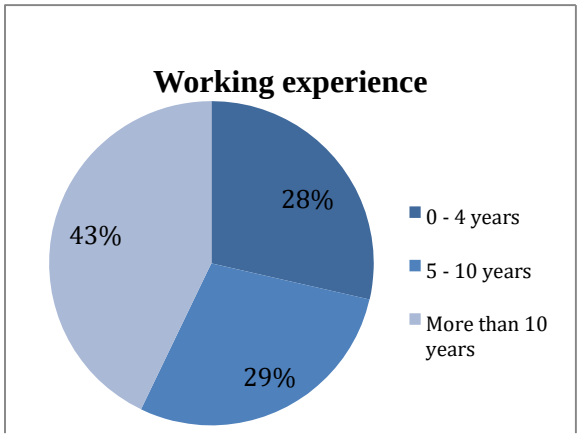


Figure 1: Respondents' working experience.

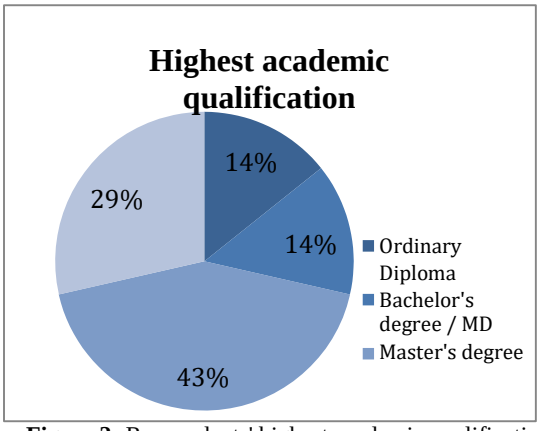


Figure 2: Respondents' highest academic qualification.

3.2 Needed Patients' Particulars for Enhanced Epidemiology

During interviews, respondents (with exception of the system administrators) were provided with a list of patients' particulars collected by the GoT-HoMIS when a patient is registered at the Hospital for the first time. The particulars included first name, middle name, surname, gender, date of birth, age, mobile number, tribe, residence, marital status, occupation, country, and next of kin (name, residence, relationship, phone number). The respondents were asked to select particulars out of the list and suggest any other particulars (demographic and clinical) that can be used to enhance epidemiology and epidemics surveillance. Out of the list, 100% (n = 7) of the respondents recommended the inclusion of gender, age, and residence as parameters in epidemics surveillance. Furthermore, 85.71% (n = 6) of the respondents voted for occupation, 57.14% (n = 4) for diagnosis, and 42.86% (n = 3) for both tribe and marital status. In parallel, 14.29% (n = 1) suggested religion despite it not being in the list. The ones in support of tribe and religion associated them with some cultural practices that may lead to transmission of diseases. Figure 3 is a visualization of the suggested attributes along with the number of respondents in support of each:

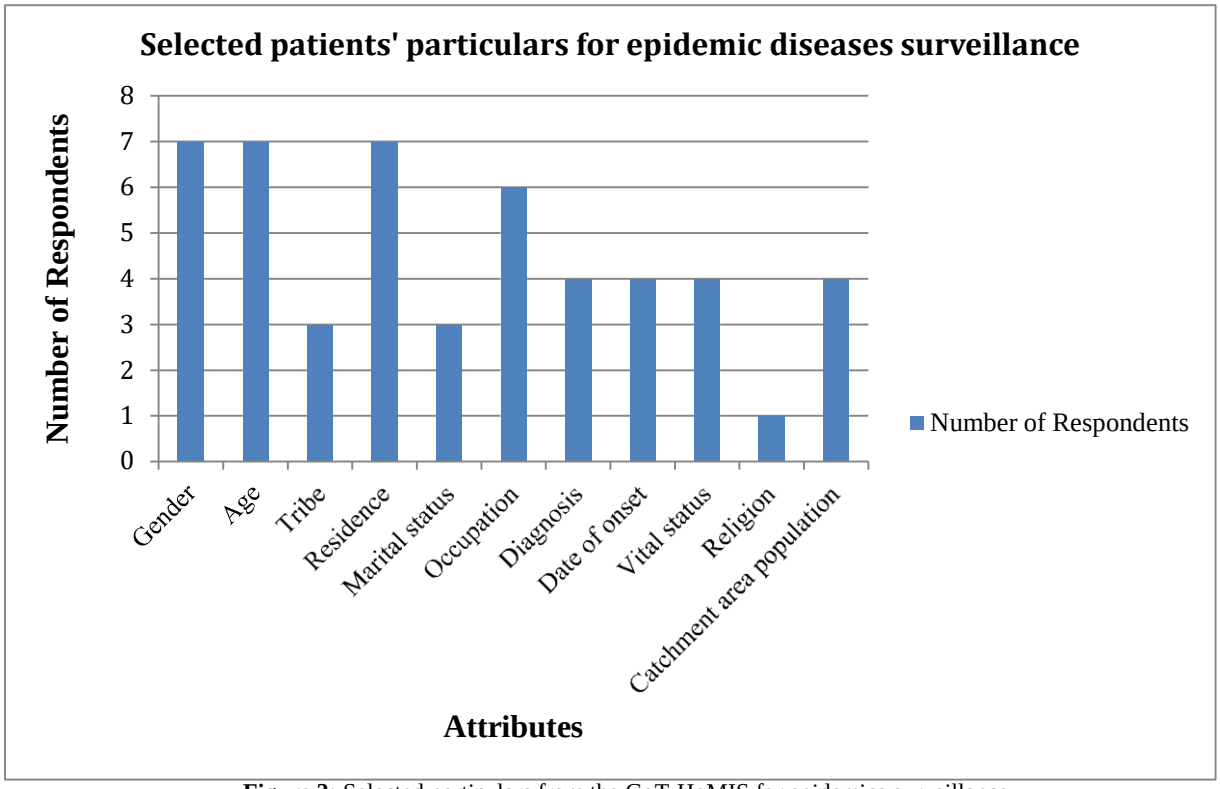


Figure 3: Selected particulars from the GoT-HoMIS for epidemics surveillance.

Catchment area population and vital status, though not included in the list provided to the respondents, were each proposed by four of the respondents (57.14%). They further explained that the

two attributes would be useful to provide insights on the severity of the case and enhance informed decision making on mitigation strategies. The catchment area population and the vital status would give insights to the attack rate (AT) and case fatality rate (CFR), respectively. For example, reporting fifty (50) cases of a disease in a population of two hundred (200) people is quite different from fifty (50) cases in a population of five thousand (5000) people. Moreover, five cases of a disease being confirmed at a hospital and all five-people dying within a short time is different from five cases of a disease where all or a significant number of the affected are still alive past the first week. Vital status was also closely associated with date of onset, which also was proposed by 57.14% (n = 4) of the respondents.

3.3 The Current Reporting Setup at the MoHCDGEC

The respondents team at the MoHCDGEC were further interviewed on the reporting process and diseases surveillance. The target was to understand the current setup and the role of the existing systems and innovations in facilitating timeliness, completeness, and enhancement of the reports. The current model derived from interviews responses is as presented in Fig. 4:

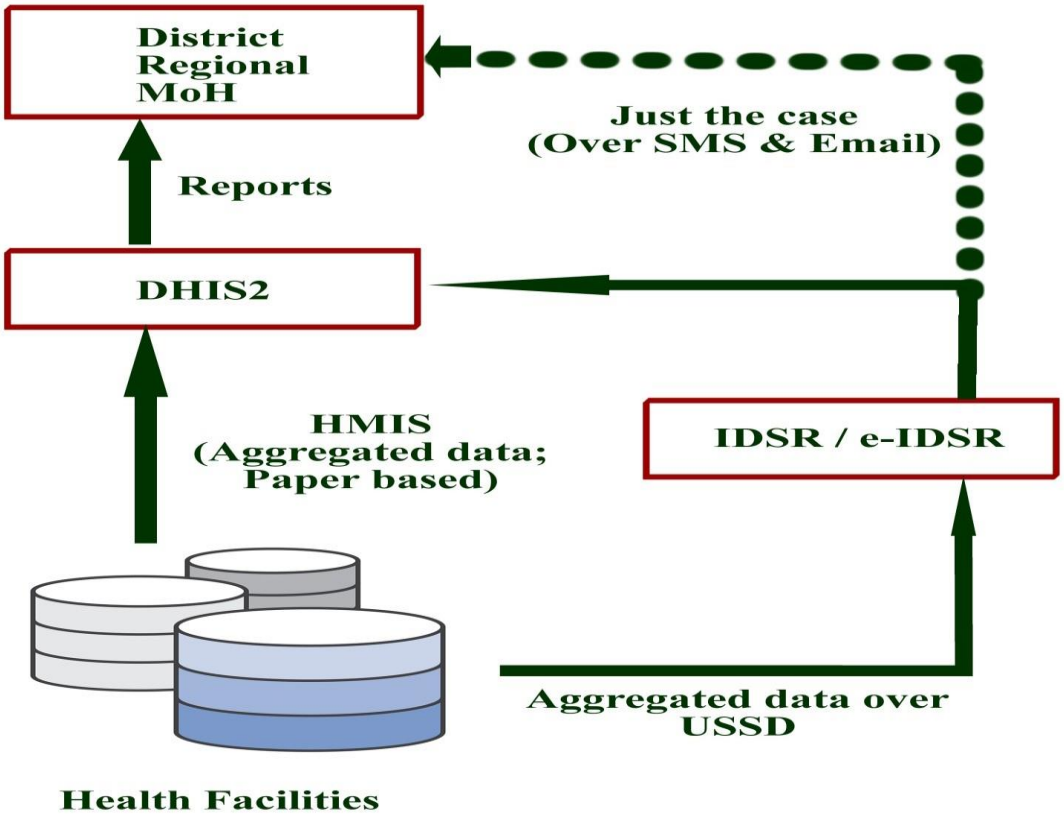


Figure 4: Current setup for disease cases reporting.

It was revealed in the interviews, as presented in Fig. 4, that the health facilities are the primary sources of information. These capture information about patients and their visits at the facilities. Different hospitals operate different Hospital Management Information Systems such as the GoT-HoMIS, Jeeva and Care2x; while, some are still paper based. Harmonization of the collected data is done through the Health Management Information System (HMIS/MTUHA), paper-based registers that contain aggregated reports on different aspects regarding clinical matters. There are a total of 16 registers, each having its own distinctive purpose and information. The patients' clinical data from the health facilities through HMIS reports are inserted in the District Health Information System (DHIS2). This operation is done on a monthly basis. Moreover, the data inserted in the DHIS2 is already in aggregated manner. The DHIS2 generates various reports that are circulated to the district, regional, and national health teams. Despite the quality of reports generated by the DHIS2, nature of the primary data entry system (from hard copy registers) and being on monthly basis results in challenges of timeliness, correctness, and completeness of reports. Some disease cases require immediate alert to epidemiologists and time-to-time follow-up, which is impossible with the DHIS2. In addition to the shortcomings, it was reported that there are often mismatches between the aggregated data presented by the DHIS2 and the data present in the health facilities.

The respondents further explained that in response to the need for timeliness of reporting and alerts for some diseases, especially outbreaks, the MoHCDGEC developed the Integrated Disease Surveillance and Response (IDSR) and its more enhanced innovation e-IDSR. These were developed to simultaneously feed data to the DHIS2 and alert district, regional, and national responsible personnel in case of an outbreak. The e-IDSR utilizes Unstructured Supplementary Service Data (USSD) to pass information to the DHIS2 as well as targeted epidemiologists. Selected people at the health facilities provide information to the system, and alerts are immediately forwarded to the epidemiologists in forms of emails and text messages. The alerts however, contain just the case (diagnosed disease), the name of the health facility where the diagnosis has been confirmed, and phone number of the contact person at the health facility for the sake of follow-up. Figure 5 shows a sample text message sent to epidemiologists.

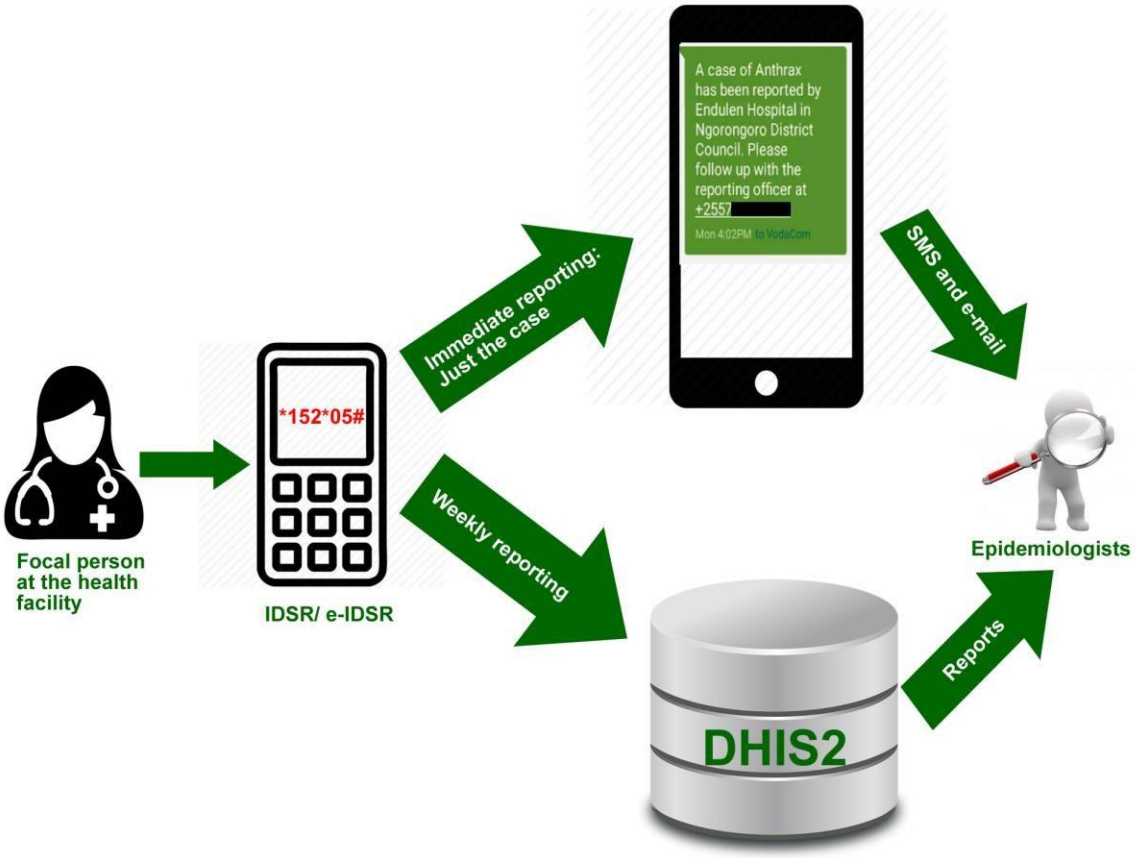


Figure 5: e-IDSR framework [11].

The epidemiologists also pointed out that this has proven to be insufficient information to work with in terms of analysis and looking for patterns in the outbreak. On a few selected diseases (epidemic prone diseases), weekly follow-up is done by feeding data into the IDSR (in aggregated manner), which consequently feeds them into the DHIS2. Reports can then be generated from the DHIS2 grouping the aggregated number of cases gender and age wise as seen in the sample report format in Fig. 6.

FORM 3 C: WEEKLY REPORTED NEW CASES / DEATHS DURING AN EPIDEMIC AT REGION LEVEL

Region: Week beginning: Week ending:

S/N	DISEASES	< 5				> 5			
		C		D		C		D	
		M	F	M	F	M	F	M	F
1	AFP								
2	Anthrax								
3	Blood Diarrhea								
4	Cholera								
5	CSM								
6	Human Influenza/SARI								
7	Keratoconjuctivitis								
8	Measles								

Figure 6: Template for an aggregated report generated by the DHIS2 [11].

One of the challenges facing e-IDSR usage is the USSD not being efficient in times of outbreaks and in large hospitals where there is a large number of disease cases. USSD are also subject to timeouts when there is delay in the feeding process. Moreover, it was documented that the MoHCDGEC must pay a monthly fee for the aggregator that collects and aggregates data from the USSDs and feeds them to the DHIS2 [11]. The e-IDSR is currently operational in 15 regions in the URT. Furthermore, three (3) respondents pointed out that data and reports generated from the health facilities only reflect a fraction of health issues and disease cases present in the country. Disease cases in the community may not be captured unless the sufferer reports to a health facility, which is not always the case. The respondents further requested for a way that community data (disease cases) can be captured and integrated in the reports for better understanding and planning of diseases mitigation strategies. Moreover, the epidemiologists mentioned rumors as one of the main sources of their information of outbreaks. Due to the advancement of technology, internet connectivity, and smartphones ownership, most rumors are aired on social media platforms. It was further stated that a significant number of the received rumors prove to be valid, and they reach the epidemiologists through their connections quicker that reports from hospitals.

3.4 Proposed Data Integration Framework

The findings of the survey pointed out a need for a data integration framework that would guarantee inclusion of data from various sources (identified patients' particulars from the GoT-HoMIS and related data from other HoMIS in the country, social media, meteorological data). A framework has

been envisioned to guide the integration of data from the diversified sources, joint analytics, and timely as well as detailed presentation of reports to enhance epidemic diseases surveillance and informed decision-making. The framework targets prompt analytics and reporting to support the Public Health Emergency Control Centers (PHOEC) as emphasized by the World Health Organization [12]. Figure 7 illustrates the proposed Data Integration Framework.

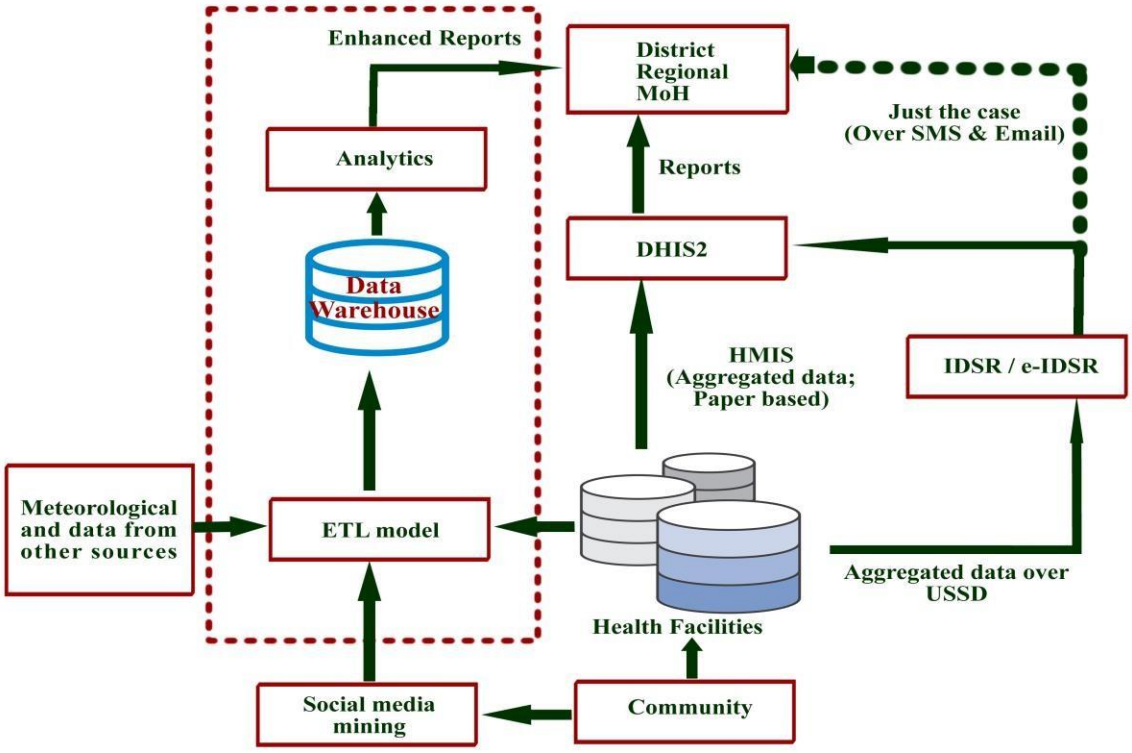


Figure 7: Proposed Data Integration Framework.

The data integration framework complements the existing reporting setup illustrated in Fig. 4. The extraction, transformation, and loading (ETL) model is targeted to fetch data from the HoMISs in the health facilities (GoT-HoMIS, Jeeva, Care2x), transform them accordingly and load them into a data warehouse for joint analytics. To enhance analytics and epidemic diseases surveillance, the ETL model will also extract and transform data from social media and load it to the data warehouse. To some extent, inclusion of means to jointly analyze rumors and epidemic diseases related posts from social media will mitigate the problem of missing analysis of community data for sufferers who do not visit health facilities.

Prompt alerts to the epidemiological teams based on timely extraction of rumors from social media will also increase effectiveness in decision making and controlling the case(s). Addition of meteorological data for integrated analysis will enable observation of patterns of disease cases in relation to the weather or climate changes.

All the aforementioned joint analytics are missing in the existing setup at the MoHCDGEC. The diversification of data sources will increase the possibility of patterns observation and extensive analytics, and consequently enhance epidemic diseases surveillance. On the resulting dataset, superior to the existing, data mining and machine learning algorithms can be applied to produce comprehensive reports and prediction of outbreaks for effective control of the epidemics. The framework does not obliterate the existing setup and innovations; the e-IDSR should still send alerts to the epidemiological teams at the district, regional, and national levels. The alerts, however, should be linked to the associated comprehensive reports that are generated from analysis on the integrated dataset.

4. Discussion

Two objectives were targeted in this research paper. The first objective was to explore the suitability of the GoT-HoMIS for epidemic diseases surveillance through the use of stored patients' particulars; and the second objective was to evaluate the current setup for epidemic diseases surveillance at the MoHCDGEC in Tanzania. From the first objective, the particulars essential for the epidemic diseases surveillance were found to be the patients' gender, age, tribe, residence, marital status, occupation, diagnosis, date of onset, vital status, religion, and catchment area population. It was established that 81.82% of the identified particulars are currently being captured and stored in the GoT-HoMIS. The exceptions are religion and catchment area population. From the identified importance of the patients' religion with ties to disease-prone practices, just as culture, the GoT-HoMIS developers should be instructed to add the field in the patients' registration form. The catchment area population can be fetched from the national census data and updated accordingly. The GoT-HoMIS can therefore be judged as a potential data source for necessary analytics in support of epidemic diseases surveillance and control.

The second objective was also successfully achieved as the current reporting setup at the MoHCDGEC has been analysed. From the health facilities, the HMIS has been identified as the bedrock for all disease cases analysis and is the primary data source to the DHIS2. While it has national-wide coverage and varieties of aggregated facility-level data, it is error-prone and still struggles with timeliness and completeness. Moreover, its frequency is not efficient for the targeted level of analysis for real-time surveillance. The MoHCDGEC's innovation for the adoption of the e-IDSR is commendable. It has successfully solved the gap of timely alerts to response teams at district, regional and national levels in case of outbreaks. Through the utilization of USSD technology it also manages to serve in hospitals with no electronic HoMIS or with poor Information and Communication Technology (ICT) infrastructure. However, reporting just the case as shown in Fig. 5 has proved ineffective to the epidemiologists. More parameters as identified in the Fig. 3 and enhanced analytics should accompany the alerts to facilitate the surveillance and informed decision-making. Furthermore, the aggregation of data before being inserted

into the DHIS2 makes it less effective for epidemic diseases surveillance. Only limited analytics can be extracted from the hosted data.

The course of integrating community data in health-related analytics as recommended by some respondents is genuine and paramount. Reporting trends of a disease case based solely on the number of people who visit health facilities is partial; there are other cases in the community affecting people who cannot or will not visit health facilities. Social media can be viewed as a potential data source for the community data, whilst working on further innovations. Fortunately, the number of Tanzanian's accessing the Internet through their mobile phones from the year 2012 has increased more than threefold, from about six (6) million people in 2012 to more than nineteen (19) million in December 2017 [13]. This increase has a ripple effect on social media subscription and usage. The posts on social media can be mined to obtain epidemic diseases related posts as well as those of epidemic-prone environments [14]. A good example of a case where the spread of an epidemic was widely posted on social media is the case of Ebola, in which countless posts were being shared on social media [15, 16]. Once the posts that contain targeted phrases or mere mention of epidemics have been extracted, they can be forwarded to epidemiologists timely, and be utilized for epidemic diseases surveillance. If well designed, social media mining can be a great contribution to epidemic diseases surveillance as well as other health related issues [15, 17, 18]. Moreover, people may post about health endangering issues such as spill of chemicals or leakage of sewers, and these can be captured and shared to epidemiologists for action. The reports from the extracted information can be analyzed along with those from hospital records and enhanced reports made available to epidemiologists.

A Data Integration Framework with an enhanced data warehouse as the nucleus for integrated analytics to support epidemic diseases surveillance is proposed (Fig. 7). Data needs to be captured from all the identified sources, that is HoMISs, social media and additional ones such as meteorological data as well as catchment area population (see Fig. 7). The data warehouse should be designed to host all the data and formulate a unified dataset that can be used for analytics and epidemics surveillance. The data loaded in the proposed warehouse should be transformed (on individual patient-wise) to allow extreme data mining operations. Aggregation will be done within the data warehouse as part of analysis should the need arise, and not before. Extraction of raw data from the sources will also reduce the mismatch in reports, contrary to what is currently experienced by the DHIS2 from the reports present on the health facilities. For the hospitals running electronic HoMIS, a new module needs to be introduced for extracting the identified parameters in Fig. 3, and loading them to the staging area, where they will be transformed and loaded into the proposed data warehouse. In case of an outbreak, the epidemiologists can still receive a notification as they do now from the e-IDSR, but through the proposed framework they can find timely system-generated analysis and determined patterns among new and historic cases.

The elements of quality data can be argued to include availability, cleanliness, completeness, correctness, and timeliness of the data. From observation done on patients' registration process, the data entrants in the systems lack knowledge on the final applications of the data they feed. Consequently, every now and then a patient is registered with incomplete particulars such as occupation and marital status, which have been identified to be crucial in the monitoring of spread of epidemic diseases. The findings support the recommendation by WHO [12], and Kimaro and Twaakyondo [19] for the need to educate all the health workers who feed patients' demographic and clinical data in the HoMISs on the essence of quality data for analysis and reports generation.

5. Conclusion and Recommendation

The study has identified the parameters to be extracted from Hospital Management Information Systems needed for epidemic diseases surveillance. The GoT-HoMIS has the potential to realize the targeted level of surveillance as it records 81.82% of the needed parameters, has large market share, and is government-owned. While existing systems at the MoHCDGEC for epidemic diseases surveillance have prospective strengths, several shortcomings have been identified to include insufficient analytics to support timely epidemic diseases surveillance, limited data mining, poor quality of data fed in the system, and missing mechanisms to capture community data. To counter the shortcomings, this paper has proposed a data integration framework to guide research and development of interventions to enable applications of machine learning and data mining in epidemic diseases surveillance with a case study of the United Republic of Tanzania's health sector. The paper also recommends educating the personnel involved in feeding data in the systems on the essence of quality data and the applications of data for analysis, as this may help improve the quality of data in the systems.

6. Conflict of Interest

The authors declare no conflict of interest.

7. Acknowledgement

Sincere appreciation is extended to the Project 2 (P2) of the VLIR-UOS GRE@T programme at Mzumbe University for funding this research. Furthermore, we wish to thank all the respondents that participated in the study together with the institutions with which they are affiliated. Lastly, gratitude is registered to the Nelson Mandela African Institution of Science and Technology for the support throughout the work.

8. References

1. Karuri J, Waiganjo P, Daniel OR, Manya A. DHIS2: the tool to improve health data demand and use in Kenya. *Journal of Health Informatics in Developing Countrie*s. 2014 Mar 18;8(1).
2. President's Office, Regional Administration and Local Government (PO-RALG). Government of Tanzania – Hospital Management Information System (GoT-HoMIS) [PowerPoint Slides]. 2017 [Cited 6th May 2018]. Available from: https://www.healthdatacollaborative.org/fileadmin/uploads/hdc/Documents/Country_documents/Tanzania_GOT-HOMIS_presentation_12Sept2017.pdf
3. Nyasulu P, Kasubi M, Boniface R, Murray J. Understanding Laboratory Methods and Their Impact on Antimicrobial Resistance Surveillance, at Muhimbili National Hospital, Dar es Salaam, Tanzania. *Advances in Microbiology*. 2014; 4:33-8.
4. Wambura W, Machuve D, Nykänen P. Development of Discharge Letter Module onto a Hospital Information System. *Journal of Health Informatics in Developing Countries*. 2017 Dec 10;11(2).
5. Sheikh YH, Nyella E. Exploring the Power of Technology in the Institutionalization of Health Information Systems: An Actor-Network Analysis of Information Systems Integration. *Journal of Health Informatics in Africa*. 2018 Jan 21;4(2).
6. Parvathi I, Rautaray S. Survey on data mining techniques for the diagnosis of diseases in medical domain. *International Journal of Computer Science and Information Technologies*. 2014;5(1):838-46.
7. Molloy GJ, Stamatakis E, Randall G, Hamer M. Marital status, gender and cardiovascular mortality: behavioural, psychological distress and metabolic explanations. *Social science & medicine*. 2009 Jul 1;69(2):223-8.
8. National Bureau of Statistics (NBS), Office of the Chief Government Statistician (OCGS, Zanzibar), Ministry of Health, Community, Development, Gender, Elderly and Children (MoHCDGEC), Ministry of Health (MoH) Zanzibar. Tanzania Malaria Indicator Survey (TMIS) Key Indicators 2017. 2018 [cited 5th May 2018]. Available from: https://nbs.go.tz/nbs/takwimu/TMIS2017/TMIS2017_KeyIndicatorEng.pdf
9. Darcy NM, Somi G, Matee M, Wengaa D, Perera S. Analysis of Data Dissemination and Use Practices in the Health Sec-tor in Tanzania: Results of desk review and interviews with key stakeholders. *Journal of Health Informatics in Africa*. 2017 Nov 5;4(1).
10. Health Metrics Network (HMN). Framework and Standards for Country Health Information Systems. Geneva, World Health Organisation 2008.
11. Ministry of Health, Community Development, Gender, Elderly and Children (MoHCDGEC). Tanzania e-IDSR System [PowerPoint Slides]. 2018.

12. World Health Organization (WHO). WHO Tanzania Annual Report 2016. 2016 [cited 8th May, 2018]. Available from: http://www.afro.who.int/publications/who-country-office-annual-report-2016
13. Tanzania Communications Regulatory Authority (TCRA). Quarterly Communications Statistics Report October- December 2017 Operator Returns. 2018 [cited 5th May, 2018]. Available from: https://www.tcra.go.tz/images/documents/telecommunication/TelCom_Statistics_Dec_2017.pdf
14. Corley CD, Cook DJ, Mikler AR, Singh KP. Text and structural data mining of influenza mentions in web and social media. *International journal of environmental research and public health*. 2010 Feb 22;7(2):596-615.
15. Fernández-Luque L, Bau T. Health and social media: perfect storm of information. *Healthcare informatics research.* 2015 Apr 1;21(2):67-73.
16. Househ M. Communicating Ebola through social media and electronic news media outlets: A cross-sectional study. *Health informatics journal*. 2016 Sep;22(3):470-8.
17. Nikfarjam A, Sarker A, O'connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the *American Medical Informatics Association.* 2015 Mar 9;22(3):671-81.
18. Paul MJ, Sarker A, Brownstein JS, Nikfarjam A, Scotch M, Smith KL, Gonzalez G. Social media mining for public health monitoring and surveillance. *InBiocomputing* 2016: Proceedings of the Pacific Symposium 2016 (pp. 468-479).
19. Kimaro HC, Twaakyondo HM. Analysing the hindrance to the use of information and technology for improving efficiency of health care delivery system in Tanzania. Tanzania *Journal of Health Research.* 2005;7(3):189-97.